EVALUATING PUBLIC PROGRAMS WITH CLOSE SUBSTITUTES: THE CASE OF HEAD START*

PATRICK KLINE AND

CHRISTOPHER R. WALTERS

We use data from the Head Start Impact Study (HSIS) to evaluate the cost-effectiveness of Head Start, the largest early childhood education program in the United States. Head Start draws roughly a third of its participants from competing preschool programs, many of which receive public funds. We show that accounting for the fiscal impacts of such program substitution pushes estimates of Head Start's benefit-cost ratio well above one under a wide range of assumptions on the structure of the market for preschool services and the dollar value of test score gains. To parse the program's test score impacts relative to home care and competing preschools, we selection-correct test scores in each care environment using excluded interactions between experimental assignments and household characteristics. We find that Head Start generates larger test score gains for children who would not otherwise attend preschool and for children who are less likely to participate in the program. *JEL* Codes: I20, J24, H52, C30.

I. INTRODUCTION

Many government programs provide services that can be obtained in roughly comparable form via markets or through other public organizations. The presence of close program substitutes complicates the task of program evaluation by generating ambiguity regarding which causal estimands are of interest. Standard intent-to-treat impacts from experimental demonstrations can yield unduly negative assessments of program effectiveness if most participants would receive similar services in the absence of an intervention (Heckman et al. 2000). On the other hand, experiments that artificially restrict substitution alternatives may yield impacts that are not representative of the costs and benefits of actual policy changes.

*We thank Danny Yagan, James Heckman, Nathan Hendren, Magne Mogstad, Jesse Rothstein, Melissa Tartari, and seminar participants at UC Berkeley, the University of Chicago, Arizona State University, Harvard University, Stanford University, the NBER Public Economics Spring 2015 Meetings, Columbia University, UC San Diego, Princeton University, the NBER Labor Studies Summer Institute 2015 Meetings, Uppsala University, MIT, Vanderbilt University, and four anonymous referees for helpful comments. Raffaele Saggio provided outstanding research assistance. We also thank Research Connections for providing the data. Generous funding support for this project was provided by the Berkeley Center for Equitable Growth.

The Quarterly Journal of Economics (2016), 1795–1848. doi:10.1093/qje/qjw027.
Advance Access publication on July 11, 2016.

This article assesses the cost-effectiveness of Head Start—a prominent public education program for which close public and private substitutes are widely available. Head Start is the largest early childhood education program in the United States. Launched in 1965 as part of President Lyndon Johnson's war on poverty, the program has evolved from an eight-week summer program into a year-round program that offers education, health, and nutrition services to disadvantaged children and their families. By 2013, Head Start enrolled about 900,000 three- and four-year-old children at a cost of $7.6 billion (U.S. DHHS 2013).

Views on the effectiveness of Head Start vary widely (Ludwig and Phillips 2007 and Gibbs, Ludwig, and Miller 2011 provide reviews). A number of observational studies find substantial short- and long-run impacts on test scores and other outcomes (Currie and Thomas 1995; Garces, Thomas, and Currie 2002; Ludwig and Miller 2007; Deming 2009; Carneiro and Ginja 2014). By contrast, a recent randomized evaluation—the Head Start Impact Study (HSIS)—finds small impacts on test scores that fade out quickly (Puma et al. 2010; Puma, Bell, and Heid 2012). These results have generally been interpreted as evidence that Head Start is ineffective and in need of reform (Barnett 2011; Klein 2011).

Two observations suggest that such conclusions are premature. First, research on early childhood interventions finds long-run gains in adult outcomes despite short-run fade-out of test score impacts (Heckman et al. 2010; Heckman, Pinto, and Savelyev 2013; Chetty et al. 2011; Chetty, Friedman, and Rockoff 2014b). Second, roughly one third of the HSIS control group participated in alternate forms of preschool. This suggests that the HSIS may have shifted many students between different sorts of preschools without altering their exposure to preschool services. The aim of this article is to clarify how the presence of substitute preschools affects the interpretation of the HSIS results and the cost-effectiveness of the Head Start program.

Our study begins by revisiting the experimental impacts of the HSIS on student test scores. We replicate the fade-out pattern found in previous work but find that adjusting for experimental non compliance leads to imprecise estimates of the effect of Head Start participation beyond the first year of the experiment. As a result, the conclusion of complete effect fade-out is less clear than naive intent-to-treat estimates suggest. Turning to substitution

patterns, we find that roughly one third of Head Start "compliers" (Angrist, Imbens, and Rubin 1996) in the HSIS experiment would have participated in other forms of preschool had they not been lotteried into the program. These alternative preschools draw heavily on public funding, which mitigates the net costs to government of shifting children from other preschools into Head Start.

These facts motivate a theoretical analysis clarifying which parameters are (and are not) policy relevant when publicly subsidized program substitutes are present. We work with a stylized model where test score impacts are valued according to their effects on children's after-tax lifetime earnings. We show that when competing preschool programs are not rationed, the policy-relevant causal parameter governing the benefits of Head Start expansion is an average effect of Head Start participation relative to the next best alternative, regardless of whether that alternative is a competing program or home care. This parameter coincides with the local average treatment effect (LATE) identified by a randomized experiment with imperfect compliance when the experiment contains a representative sample of program compliers. Hence, imperfect compliance and program substitution, often thought to be confounding limitations of social experiments, turn out to be virtues when the substitution patterns in the experiment replicate those found in the broader population.

We use this result to derive an estimable benefit-cost ratio associated with Head Start expansions. This ratio scales Head Start's projected impacts on the after-tax earnings of children by its net costs to government inclusive of fiscal externalities. Chief among these externalities is the cost savings that arise when Head Start draws children away from competing subsidized preschool programs. Although such effects are typically ignored in cost-benefit analyses of Head Start and other similar programs (e.g., CEA 2015), we find via a calibration exercise that such omissions can be quantitatively important: Head Start roughly breaks even when the cost savings associated with program substitution are ignored, but yields benefits nearly twice as large as costs when these savings are incorporated. This appears to be a robust finding—after accounting for fiscal externalities, Head Start's benefits exceed its costs whenever short-run test score impacts yield earnings gains within the range found in the recent literature.

A limitation of our baseline analysis is that it assumes changes in program scale do not alter the mix of program compliers. To address this issue, we also consider "structural reforms" to Head Start that change the mix of compliers without affecting test score outcomes. Examples of such reforms might include increased transportation services, marketing efforts, or spending on program features that parents value. Households who respond to structural reforms may differ from experimental compliers on unobserved dimensions, including their mix of counterfactual program choices. Assessing these reforms therefore requires knowledge of parameters not directly identified by the HSIS experiment. Specifically, we show that such reforms require identification of a variant of the marginal treatment effect (MTE) concept of Heckman and Vytlacil (1999).

To assess reforms that attract new children, we develop a selection model that parameterizes variation in treatment effects with respect to counterfactual care alternatives as well as observed and unobserved child characteristics. We prove that the model parameters are identified and propose a two-step control function estimator that exploits heterogeneity in the response to Head Start offers across sites and demographic groups to infer relationships between unobserved factors driving preschool enrollment and potential outcomes. The estimator is shown to pass a variety of specification tests and accurately reproduce patterns of treatment effect heterogeneity found in the experiment. The model estimates indicate that Head Start has large positive short-run effects on the test scores of children who would have otherwise been cared for at home and insignificant effects on children who would otherwise attend other preschools—a finding corroborated by Feller et al. (forthcoming), who reach similar conclusions using principal stratification methods (Frangakis and Rubin 2002). Our estimates also reveal a "reverse Roy" pattern of selection whereby children with unobserved characteristics that make them less likely to enroll in Head Start experience larger test score gains.

We conclude with an assessment of prospects for increasing Head Start's rate of return via outreach to new populations. Our estimates suggest that expansions of Head Start could boost the program's rate of return provided that the proposed technology for increasing enrollment (e.g., improved transportation services) is not too costly. We also use our estimated selection model to examine the robustness of our results to rationing of competing

preschools. Rationing implies that competing subsidized preschools do not contract when Head Start expands, which shuts down a form of public savings. On the other hand, expanding Head Start generates opportunities for new children to fill vacated seats in substitute programs. Our estimates indicate that the effect on test scores (and therefore earnings) of moving children from home care to competing preschools is substantial, leading us to conclude that rationing is in fact likely to increase the favorable estimated rates of return found in our baseline analysis.

The rest of the article is structured as follows. Section II provides background on Head Start. Section III describes the HSIS data and basic experimental impacts. Section IV presents evidence on substitution patterns. Section V introduces a theoretical framework for assessing public programs with close substitutes. Section VI provides a cost-benefit analysis of Head Start. Section VII develops our econometric selection model and discusses identification and estimation. Section VIII reports estimates of the model. Section IX simulates the effects of structural program reforms. Section X concludes.

II. Background on Head Start

Head Start provides preschool for disadvantaged children in the United States. The program is funded by federal grants awarded to local public or private organizations. Grantees are required to match at least 20% of their Head Start awards from other sources and must meet a set of program-wide performance criteria. Eligibility for Head Start is generally limited to children from households below the federal poverty line, though families above this threshold may be eligible if they meet other criteria, such as participation in the Temporary Aid for Needy Families (TANF) program. Up to 10% of a Head Start center's enrollment can also come from higher-income families. The program is free: Head Start grantees are prohibited from charging families fees for services (U.S. DHHS 2014). It is also oversubscribed: in 2002, 85% of Head Start participants attended programs with more applicants than available seats (Puma et al. 2010).

Head Start is not the only form of subsidized preschool available to poor families. Preschool participation rates for disadvantaged children have risen over time as cities and states expanded

their public preschool offerings (Cascio and Schanzenbach 2013). Moreover, the Child Care Development Fund program provides block grants that finance childcare subsidies for low-income families, often in the form of child care vouchers that can be used for center-based preschool (U.S. DHHS 2012). Most states also use TANF funds to finance additional child care subsidies (Schumacher, Greenberg, and Duffy 2001). Because Head Start services are provided by local organizations who themselves must raise outside funds, it is unclear to what extent Head Start and other public preschool programs actually differ in their education technology.

A large nonexperimental literature suggests that Head Start produced large short- and long-run benefits for early cohorts of program participants. Several studies estimate the effects of Head Start by comparing program participants to their nonparticipant siblings (Currie and Thomas 1995; Garces, Thomas, and Currie 2002; Deming 2009). Results from this research design show positive short-run effects on test scores and long-run effects on educational attainment, earnings, and crime. Other studies exploit discontinuities in Head Start program rules to infer program effects (Ludwig and Miller 2007; Carneiro and Ginja 2014). These studies show longer run improvements in health outcomes and criminal activity.

In contrast to these nonexperimental estimates, results from a recent randomized controlled trial reveal smaller, less persistent effects. The 1998 Head Start reauthorization bill included a congressional mandate to determine the effects of the program. This mandate resulted in the HSIS: an experiment in which more than 4,000 applicants were randomly assigned via lottery to either a treatment group with access to Head Start or a control group without access in fall 2002. The experimental results showed that a Head Start offer increased measures of cognitive achievement by roughly 0.1 standard deviations during preschool, but these gains faded out by kindergarten. Moreover, the experiment showed little evidence of effects on noncognitive or health outcomes (Puma et al. 2010; Puma, Bell, and Heid 2012). These results suggest both smaller short-run effects and faster fade-out than nonexperimental estimates for earlier cohorts. Scholars and policy makers have generally interpreted the HSIS results as evidence that Head Start is ineffective and in need of reform (Barnett 2011). The experimental results have also been cited in the popular media to motivate calls for dramatic

restructuring or elimination of the program (Klein 2011; Stossel 2014).[1]

Differences between the HSIS results and the nonexperimental literature could be due to changes in program effectiveness over time or to selection bias in nonexperimental sibling comparisons. Another explanation, however, is that these two research designs identify different parameters. Most nonexperimental analyses have focused on recovering the effect of Head Start relative to home care. In contrast, the HSIS measures the effect of Head Start relative to a mix of alternative care environments, including other preschools.

III. Data and Experimental Impacts

Before turning to an analysis of program substitution issues, we describe the HSIS data and report experimental impacts on test scores and program compliance.

III.A. Data

Our core analysis sample includes 3,571 HSIS applicants with nonmissing baseline characteristics and spring 2003 test scores. Online Appendix A describes construction of this sample. The outcome of interest is a summary index of cognitive test scores that averages Woodcock Johnson III (WJIII) test scores with Peabody Picture and Vocabulary Test (PPVT) scores, with each test normed to have mean 0 and variance 1 in the control group by cohort and year. We use WJIII and PPVT scores because these are among the most reliable tests in the HSIS data; both are available in each year of the experiment, allowing us to produce comparable estimates over time.

1. Subsequent analyses of the HSIS data suggest caveats to this negative interpretation but do not overturn the finding of modest mean test score impacts accompanied by rapid fade-out. Gelber and Isen (2013) find persistent effects on parental engagement with children. Bitler, Domina, and Hoynes (2014) find larger experimental impacts at low quantiles of the test score distribution. These quantile treatment effects fade out by first grade, though there is some evidence of persistent effects at the bottom of the distribution for Spanish speakers. Walters (2015) finds evidence of substantial heterogeneity in impacts across experimental sites and investigates the relationship between this heterogeneity and observed program characteristics. Walters finds smaller effects for Head Start centers that draw more children from other preschools rather than home care, a finding we explore in more detail here.

Table I provides summary statistics for our analysis sample. The HSIS experiment included two age cohorts: 55% of applicants were randomized at age three and could attend Head Start for up to two years, while the remaining 45% were randomized at age four and could attend for up to one year. The demographic information in Table I shows that the Head Start population is disadvantaged. Less than half of Head Start applicants live in two-parent households, and the average applicant's household earns about 90% of the federal poverty line. Column (2) of Table I compares these and other baseline characteristics for the HSIS treatment and control groups to check balance in randomization. The results here indicate that randomization was successful: baseline characteristics were similar for offered and non offered applicants.[2]

Columns (3) through (5) of Table I report summary statistics for children attending Head Start, other preschool centers, and no preschool.[3] Children in other preschools tend to be less disadvantaged than children in Head Start or no preschool, though most differences between these groups are modest. The other preschool group has a lower share of high school dropout mothers, a higher share of mothers who attended college, and higher average household income than the Head Start and no preschool groups. Children in other preschools outscore the other groups by about 0.1 standard deviations on a baseline summary index of cognitive skills. The other preschool group also includes a relatively large share of four-year-olds, likely reflecting the fact that alternative preschool options are more widely available for four-year-olds (Cascio and Schanzenbach 2013).

III.B. Experimental Impacts

Table II reports experimental impacts on test scores. Columns (1), (4), and (7) report intent-to-treat impacts of the

2. Random assignment in the HSIS occurred at the Head Start center level, and offer probabilities differed across centers. We weight all models by the inverse probability of a child's assignment, calculated as the site-specific fraction of children assigned to the treatment group.

3. Preschool attendance is measured from the HSIS "focal arrangement type" variable, which combines information from parent interviews and teacher/care provider interviews to construct a summary measure of the childcare setting. See Online Appendix A for details.

TABLE I

DESCRIPTIVE STATISTICS

	By offer status			By preschool choice		
Variable	(1) Offered mean	(2) Nonoffered mean	(3) Differential	(4) Head Start	(5) Other centers	(6) No preschool
Male	0.494	0.505	−0.011	0.501	0.506	0.492
			(0.019)			
Black	0.308	0.298	0.010	0.317	0.353	0.250
			(0.010)			
Hispanic	0.376	0.369	0.007	0.380	0.354	0.373
			(0.010)			
Teen mother	0.159	0.174	−0.015	0.159	0.169	0.176
			(0.014)			
Mother married	0.436	0.448	−0.011	0.439	0.420	0.460
			(0.017)			
Both parents in household	0.497	0.488	0.009	0.497	0.468	0.499
			(0.017)			
Mother is high school dropout	0.368	0.397	−0.029	0.377	0.322	0.426
			(0.017)			
Mother attended some college	0.298	0.281	0.017	0.293	0.342	0.253
			(0.016)			
Spanish speaker	0.287	0.273	0.014	0.296	0.274	0.260
			(0.011)			
Special education	0.136	0.108	0.028	0.134	0.145	0.091
			(0.011)			

TABLE I
(CONTINUED)

	By offer status			By preschool choice		
Variable	(1) Offered mean	(2) Nonoffered mean	(3) Differential	(4) Head Start	(5) Other centers	(6) No preschool
Only child	0.161	0.139	0.022	0.151	0.190	0.123
			(0.012)			
Income (fraction of FPL)	0.896	0.896	0.000	0.892	0.983	0.851
			(0.024)			
Age 4 cohort	0.448	0.451	−0.003	0.426	0.567	0.413
			(0.012)			
Baseline summary index	0.003	0.012	−0.009	−0.001	0.106	−0.040
			(0.027)			
Urban	0.833	0.835	−0.002	0.834	0.859	0.819
			(0.003)			
Center provides transportation	0.606	0.604	0.002	0.586	0.614	0.628
			(0.005)			
Center quality index	0.465	0.470	−0.005	0.452	0.474	0.488
			(0.005)			
Joint p-value			0.506			
N	2,256	1,315	3,571	2,043	598	930

Notes. All statistics weight by the reciprocal of the probability of a child's experimental assignment. Standard errors are clustered at the center level. The transportation and quality index variables refer to a child's Head Start center of random assignment. The quality variable combines information on center characteristics (teacher and center director education and qualifications, class size) and practices (variety of literacy and math activities, home visiting, health and nutrition). Income is missing for 19% of observations. Missing values are excluded in statistics for income. The baseline summary index is the average of standardized PPVT and WJIII scores in fall 2002, with each score standardized to have mean 0 and standard deviation 1 in the control group separately by applicant cohort. The joint p-value is from a test of the hypothesis that all coefficients equal 0.

TABLE II

EXPERIMENTAL IMPACTS ON TEST SCORES

	Three-year-old cohort			Four-year-old cohort			Cohorts pooled		
Time period	(1) Reduced form	(2) First stage	(3) IV	(4) Reduced form	(5) First stage	(6) IV	(7) Reduced form	(8) First stage	(9) IV
Year 1	0.194	0.699	0.278	0.141	0.663	0.213	0.168	0.682	0.247
	(0.029)	(0.025)	(0.041)	(0.029)	(0.022)	(0.044)	(0.021)	(0.018)	(0.031)
N		1,970			1,601			3,571	
Year 2	0.087	0.356	0.245	−0.015	0.670	−0.022	0.046	0.497	0.093
	(0.029)	(0.028)	(0.080)	(0.037)	(0.023)	(0.054)	(0.024)	(0.020)	(0.049)
N		1,760			1,416			3,176	
Year 3	−0.010	0.365	−0.027	0.054	0.666	0.081	0.019	0.500	0.038
	(0.031)	(0.028)	(0.085)	(0.040)	(0.025)	(0.060)	(0.025)	(0.020)	(0.050)
N		1,659			1,336			2,995	
Year 4	0.038	0.344	0.110		—			—	
	(0.034)	(0.029)	(0.098)						
N		1,599							

Notes. This table reports experimental estimates of the effects of Head Start on test scores. The outcome is the average of standardized PPVT and WJIII scores, with each score standardized to have mean 0 and standard deviation 1 in the control group separately by applicant cohort and year. Columns (1), (4), and (7) report coefficients from regressions of test scores on an indicator for assignment to Head Start. Columns (2), (5), and (8) report coefficients from first-stage regressions of Head Start attendance on Head Start assignment. The attendance variable is an indicator equal to 1 if a child attends Head Start at any time prior to the test. Columns (3), (6), and (9) report coefficients from two-stage least squares (2SLS) models that instrument Head Start attendance with Head Start assignment. All models weight by the reciprocal of a child's experimental assignment, and control for sex, race, Spanish language, teen mother, mother's marital status, presence of both parents in the home, family size, special education status, income quartile dummies, urban, and a cubic polynomial in baseline score. Missing values for covariates are set to 0, and dummies for missing are included. Standard errors are clustered by center of random assignment.

Head Start offer, separately by year and age cohort. To increase precision, we regression-adjust these treatment/control differences using the baseline characteristics in Table I.[4] The intent-to-treat estimates mirror those previously reported in the literature (e.g., Puma et al. 2010). In the first year of the experiment, children offered Head Start scored higher on the summary index. For example, three-year-olds offered Head Start gained 0.19 standard deviations in test score outcomes relative to those denied Head Start. The corresponding effect for four-year-olds is 0.14 standard deviations. However, these gains diminish rapidly: the pooled impact falls to a statistically insignificant 0.02 standard deviations by year 3. Our data include a fourth year of follow-up for the three-year-old cohort. Here, too, the intent-to-treat is small and statistically insignificant (0.038 standard deviations).

Interpretation of these intent-to-treat impacts is clouded by noncompliance with random assignment. Columns (2), (5), and (8) of Table II report first-stage effects of assignment to Head Start on the probability of participating in Head Start, and columns (3), (6), and (9) report instrumental variables (IV) estimates, which scale the intent-to-treat estimates by the first-stage estimates.[5] These estimates can be interpreted as local average treatment effects (LATEs) for "compliers"—children who respond to the Head Start offer by enrolling in Head Start. Assignment to Head Start increases the probability of participation by two thirds in the first year after random assignment. The

4. The control vector includes sex, race, assignment cohort, teen mother, mother's education, mother's marital status, presence of both parents, an only child dummy, a Spanish language indicator, dummies for quartiles of family income and missing income, urban status, an indicator for whether the Head Start center provides transportation, an index of Head Start center quality, and a third-order polynomial in baseline test scores.

5. Here we define Head Start participation as enrollment at any time prior to the test. This definition includes attendance at Head Start centers outside the experimental sample. An experimental offer may cause some children to switch from an out-of-sample center to an experimental center; if the quality of these centers differs, the exclusion restriction required for our IV approach is violated. Online Appendix Table A.I compares characteristics of centers attended by children in the control group (always takers) to those of the experimental centers to which these children applied. These two groups of centers are very similar, suggesting that substitution between Head Start centers is unlikely to bias our estimates.

corresponding IV estimate implies that Head Start attendance boosts first-year test scores by 0.247 standard deviations.

Compliance for the three-year-old cohort falls after the first year as members of the control group reapply for Head Start, resulting in larger standard errors for estimates in later years of the experiment. The first stage for three-year-olds falls to 0.36 in the second year, whereas the intent-to-treat falls roughly in proportion, generating a second-year IV estimate of 0.245 for this cohort. Estimates in years 3 and 4 are statistically insignificant and imprecise. The fourth-year estimate for the three-year-old cohort (corresponding to first grade) is 0.110 standard deviations, with a standard error of 0.098. The corresponding first grade estimate for four-year-olds is 0.081 with a standard error of 0.060. Notably, the 95% confidence intervals for first-grade impacts include effects as large as 0.2 standard deviations for four-year-olds and 0.3 standard deviations for three-year-olds. These results show that although the longer run estimates are insignificant, they are also imprecise due to experimental noncompliance. Evidence for fade-out is therefore less definitive than the naive intent-to-treat estimates suggest. This observation helps to reconcile the HSIS results with observational studies based on sibling comparisons, which show effects that partially fade out but are still detectable in elementary school (Currie and Thomas 1995; Deming 2009).[6]

IV. Program Substitution

We turn now to documenting program substitution in the HSIS and how it influences our results. It is helpful to develop some notation to describe the role of alternative care environments. Each Head Start applicant participates in one of three possible treatments: Head Start, which we label h; other center-based

6. One might also be interested in the effects of Head Start on noncognitive outcomes, which appear to be important mediators of the effects of early childhood programs in other contexts (Chetty et al. 2011; Heckman, Pinto, and Savelyev 2013). The HSIS includes short-run parent-reported measures of behavior and teacher-reported measures of teacher–student relationships, and Head Start appears to have no impact on these outcomes (Puma et al. 2010; Walters 2015). The HSIS noncognitive outcomes differ significantly from those analyzed in previous studies, however, and it is unclear whether they capture the same skills.

preschool programs, which we label c; and no preschool (i.e., home care), which we label n. Let $Z_i \in \{0,1\}$ indicate whether household i has a Head Start offer, and $D_i(z) \in \{h,c,n\}$ denote household i's potential treatment status as a function of the offer. Then observed treatment status can be written $D_i = D_i(Z_i)$.

The structure of the HSIS leads to natural theoretical restrictions on substitution patterns. We expect a Head Start offer to induce some children who would otherwise participate in c or n to enroll in Head Start. By revealed preference, no child should switch between c and n in response to a Head Start offer, and no child should be induced by an offer to leave Head Start. These restrictions can be expressed succinctly by the following condition:

$$D_i(1) \neq D_i(0) \Rightarrow D_i(1) = h, \tag{1}$$

which extends the monotonicity assumption of Imbens and Angrist (1994) to a setting with multiple counterfactual treatments. This restriction states that anyone who changes behavior as a result of the Head Start offer does so to attend Head Start.[7]

Under restriction (1), the population of Head Start applicants can be partitioned into five groups defined by their values of $D_i(1)$ and $D_i(0)$:

(i) n-compliers: $D_i(1) = h$, $D_i(0) = n$,
(ii) c-compliers: $D_i(1) = h$, $D_i(0) = c$,
(iii) n-never takers: $D_i(1) = D_i(0) = n$,
(iv) c-never takers: $D_i(1) = D_i(0) = c$,
(v) always takers: $D_i(1) = D_i(0) = h$.

The n- and c-compliers switch to Head Start from home care and competing preschools, respectively, when offered a seat. The two groups of never takers choose not to attend Head Start regardless of the offer. Always takers manage to enroll in Head Start even when denied an offer, presumably by applying to other Head Start centers outside the HSIS sample.

Using this rubric, the group of children enrolled in alternative preschool programs is a mixture of c-never takers and c-compliers denied Head Start offers. Similarly, the group of children in

7. See Engberg et al. (2014) for discussion of related restrictions in the context of attrition from experimental data.

home care includes n-never takers and n-compliers without offers. The two complier subgroups switch into Head Start when offered admission; as a result, the set of children enrolled in Head Start is a mixture of always takers and the two groups of offered compliers.

IV.A. Substitution Patterns

Table III presents empirical evidence on substitution patterns by comparing program participation choices for offered and nonoffered households. In the first year of the experiment 8.3% of households decline Head Start offers in favor of other preschool centers; this is the share of c-never takers. Similarly, column (3) shows that 9.5% of households are n-never takers. As can be seen in column (4), 13.6% of households manage to attend Head Start without an offer, which is the share of always takers. The Head Start offer reduces the share of children in other centers from 31.5% to 8.3%, and reduces the share of children in home care from 55% to 9.5%. This implies that 23.2% of households are c-compliers and 45.5% are n-compliers.

Notably, in the first year of the experiment, three-year-olds have uniformly higher participation rates in Head Start and lower participation rates in competing centers, which likely reflects the fact that many state-provided programs only accept four-year-olds. In the second year of the experiment, participation in Head Start drops among children in the three-year-old cohort with a program offer, suggesting that many families enrolled in the first year decided that Head Start was a bad match for their child. We also see that Head Start enrollment rises among those families that did not obtain an offer in the first round, which reflects reapplication behavior.

IV.B. Interpreting IV

How do the substitution patterns displayed in Table III affect the interpretation of the HSIS test score impacts? Let $Y_i(d)$ denote child i's potential test score if he or she participates in treatment $d \in \{h,c,n\}$. Observed scores are given by $Y_i = Y_i(D_i)$. We assume that Head Start offers affect test scores only through program participation choices. Under assumption (1), IV identifies a variant of the LATE of Imbens and Angrist (1994), giving the average effect of Head Start participation for compliers

TABLE III

PRESCHOOL CHOICES BY YEAR, COHORT, AND OFFER STATUS

		Offered			Not offered			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Time period	Cohort	Head Start	Other centers	No preschool	Head Start	Other centers	No preschool	*C*-complier share
Year 1	3-year-olds	0.851	0.058	0.092	0.147	0.256	0.597	0.282
	4-year-olds	0.787	0.114	0.099	0.122	0.386	0.492	0.410
	Pooled	0.822	0.083	0.095	0.136	0.315	0.550	0.338
Year 2	3-year-olds	0.657	0.262	0.081	0.494	0.379	0.127	0.719

Notes. This table reports shares of offered and nonoffered students attending Head Start, other center-based preschools, and no preschool, separately by year and age cohort. All statistics are weighted by the reciprocal of the probability of a child's experimental assignment. Column (7) reports estimates of the share of compliers drawn from other preschools, given by minus the ratio of the offer's effect on attendance at other preschools to its effect on Head Start attendance.

relative to a mix of program alternatives. Specifically, under equation (1) and excludability of Head Start offers:

$$\begin{aligned}(2)\qquad &\frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[1\{D_i=h\}|Z_i=1]-E[1\{D_i=h\}|Z_i=0]}\\ &=E[Y_i(h)-Y_i(D_i(0))|D_i(1)=h,D_i(0)\neq h]\\ &\equiv LATE_h.\end{aligned}$$

The left-hand side of equation (2) is the population coefficient from a model that instruments Head Start attendance with the Head Start offer. This equation implies that the IV strategy employed in Table II yields the average effect of Head Start for compliers relative to their own counterfactual care choices, a quantity we label $LATE_h$.

We can decompose $LATE_h$ into a weighted average of "subLATEs" measuring the effects of Head Start for compliers drawn from specific counterfactual alternatives as follows:

$$(3)\qquad LATE_h = S_c LATE_{ch} + (1-S_c)LATE_{nh},$$

where $LATE_{dh} \equiv E[Y_i(h)-Y_i(d)|D_i(1)=h,D_i(0)=d]$ gives the average treatment effect on d-compliers for $d \in \{c,n\}$, and the weight $S_c \equiv \frac{P(D_i(1)=h,D_i(0)=c)}{P(D_i(1)=h,D_i(0)\neq h)}$ gives the fraction of compliers drawn from other preschools.

Column (7) of Table III reports estimates of S_c by year and cohort, computed as minus the ratio of the Head Start offer's effect on other preschool attendance to its effect on Head Start attendance (see Online Appendix B). In the first year of the HSIS experiment, 34% of compliers would have otherwise attended competing preschools. IV estimates combine effects for these compliers with effects for compliers who would not otherwise attend preschool.

As detailed in Online Appendix D, the competing preschools attended by c-compliers are largely publicly funded and provide services roughly comparable to Head Start. The modal alternative preschool is a state-provided preschool program, whereas others receive funding from a mix of public sources (see Online Appendix Table A.II). Moreover, it is likely that even Head Start–eligible children attending private preschool centers receive public funding (e.g., through CCDF or TANF subsidies). Next we consider the implications of substitution from these alternative preschools for assessments of Head Start's cost-effectiveness.

V. A Model of Head Start Provision

In this section, we develop a model of Head Start participation with the goal of conducting a cost-benefit analysis that acknowledges the presence of publicly subsidized program substitutes. Our model is highly stylized and focuses on obtaining an estimable lower bound on the rate of return to potential reforms of Head Start measured in terms of lifetime earnings. The analysis ignores redistributional motives and any effects of human capital investment on criminal activity (Lochner and Moretti 2004; Heckman et al. 2010), health (Deming 2009; Carneiro and Ginja 2014), or grade repetition (Currie 2001). Adding such features would tend to raise the implied return to Head Start. We also abstract from parental labor supply decisions because prior analyses of the HSIS find no short-term impacts on parents' work decisions (Puma et al. 2010).[8] Again, incorporating parental labor supply responses would likely raise the program's rate of return.

Our discussion emphasizes that the cost-effectiveness of Head Start is contingent on assumptions regarding the structure of the market for preschool services and the nature of the specific policy reforms under consideration. Building on the heterogeneous effects framework of the previous section, we derive expressions for policy relevant "sufficient statistics" (Chetty 2009) in terms of causal effects on student outcomes. Specifically, we show that a variant of the LATE concept of Imbens and Angrist (1994) is policy relevant when considering program expansions in an environment where slots in competing preschools are not rationed. With rationing, a mix of LATEs becomes relevant, which poses challenges to identification with the HSIS data. When considering reforms to Head Start program features that change selection into the program, the policy-relevant parameter is shown to be a variant of the MTE concept of Heckman and Vytlacil (1999).

V.A. Setup

Consider a population of households, indexed by i, each with a single preschool-aged child. Each household can enroll its child in Head Start, enroll in a competing preschool program

8. We replicate this analysis for our sample in Online Appendix Table A.III, which shows that a Head Start offer has no effect on the probability that a child's mother works or on the likelihood of working full- versus part-time. Recent work by Long (2015) suggests that Head Start may have small effects on full- versus part-time work for mothers of three-year-olds.

(e.g., state-subsidized preschool), or care for the child at home. The government rations Head Start participation via program offers Z_i, which arrive at random via lottery with probability $\delta \equiv P(Z_i = 1)$. Offers are distributed in a first period. In a second period, households make enrollment decisions. Tenacious applicants who have not received an offer can enroll in Head Start by exerting additional effort. We begin by assuming that competing programs are not rationed and then relax this assumption later.

Each household has utility over its enrollment options given by the function $U_i(d,z)$. The argument $d \in \{h,c,n\}$ indexes child care environments, and the argument $z \in \{0, 1\}$ indexes offer status. Head Start offers raise the value of Head Start and have no effect on the value of other options, so that:

$$U_i(h,1) > U_i(h,0), U_i(c,z) = U_i(c), U_i(n,z) = U_i(n).$$

Household i's enrollment choice, as a function of its offer status z, is given by:

$$D_i(z) = \arg\max_{d \in \{h,c,n\}} U_i(d,z).$$

It is straightforward to show that this model satisfies the monotonicity restriction (1). Because offers are assigned at random, market shares for the three care environments can be written $P(D_i = d) = \delta P(D_i(1) = d) + (1 - \delta)P(D_i(0) = d)$.

V.B. *Benefits and Costs*

Debate over the effectiveness of educational programs often centers on their test score impacts. A standard means of valuing such impacts is in terms of their effects on later life earnings (Heckman et al. 2010; Chetty et al. 2011; Heckman, Pinto, and Savelyev 2013; Chetty, Friedman, and Rockoff 2014b).[9] Let the symbol B denote the total after-tax lifetime income of a cohort of children. We assume that B is linked to test scores by the equation:

$$B = B_0 + (1 - \tau)pE[Y_i], \tag{4}$$

where p gives the market price of human capital, τ is the tax rate faced by the children of eligible households, and B_0 is an intercept reflecting how test scores are normed. Our focus on

9. Online Appendix C considers how such valuations should be adjusted when test score impacts yield labor supply responses.

mean test scores neglects distributional concerns that may lead us to undervalue Head Start's test score impacts (see Bitler, Domina, and Hoynes 2014).

The net costs to government of financing preschool are given by:

$$C = C_0 + \varphi_h P(D_i = h) + \varphi_c P(D_i = c) - \tau p E[Y_i], \tag{5}$$

where the term C_0 reflects fixed costs of administering the program and φ_h gives the administrative cost of providing Head Start services to an additional child. Likewise, φ_c gives the administrative cost to government of providing competing preschool services (which often receive public subsidies) to another student. The term $\tau p E[Y_i]$ captures the revenue generated by taxes on the adult earnings of Head Start–eligible children. This formulation abstracts from the fact that program outlays must be determined before the children enter the labor market and begin paying taxes, a complication we adjust for in our empirical work via discounting.

V.C. Changing Offer Probabilities

Consider the effects of adjusting Head Start enrollment by changing the rationing probability δ. An increase in δ draws additional households into Head Start from competing programs and home care. As shown in Online Appendix C, the effect of a change in the offer rate δ on average test scores is given by:

$$\frac{\partial E[Y_i]}{\partial \delta} = \underbrace{LATE_h}_{\text{Effect on compliers}} \times \underbrace{\frac{\partial P(D_i = h)}{\partial \delta}}_{\text{Complier density}}. \tag{6}$$

In words, the aggregate impact on test scores of a small increase in the offer rate equals the average impact of Head Start on complier test scores times the measure of compliers. By the arguments in Section IV, both $LATE_h$ and $\frac{\partial}{\partial \delta} P(D_i = h) = P(D_i(1) = h, D_i(0) \neq h)$ are identified by the HSIS experiment. Hence, equation (6) implies that the hypothetical effects of a market-level change in offer probabilities can be inferred from an individual-level randomized trial with a fixed offer probability. This convenient result follows from the assumption that Head Start offers are distributed at random and that δ does not directly enter the alternative specific choice utilities, which in turn implies that the composition of compliers (and hence $LATE_h$) does not change with δ. Later we explore how

this expression changes when the composition of compliers responds to a policy change.

From equation (4), the marginal benefit of an increase in δ is given by:

$$\frac{\partial B}{\partial \delta} = (1-\tau) pLATE_h \times \frac{\partial P(D_i = h)}{\partial \delta}.$$

The offsetting marginal cost to government of financing such an expansion can be written:

$$(7)\quad \frac{\partial C}{\partial \delta} = \left(\underbrace{\varphi_h}_{\text{Provision Cost}} - \underbrace{\varphi_c S_c}_{\text{Public Savings}} - \underbrace{\tau pLATE_h}_{\text{Added Revenue}} \right) \times \frac{\partial P(D_i = h)}{\partial \delta}.$$

This cost consists of the measure of compliers times the administrative cost φ_h of enrolling them in Head Start minus the probability S_c that a complying household comes from a substitute preschool times the expected government savings φ_c associated with reduced enrollment in substitute preschools. The quantity $\varphi_h - \varphi_c S_c$ can be viewed as a LATE of Head Start on government spending for compliers. Subtracted from this effect is any extra revenue the government gets from raising the productivity of the children of complying households.

The ratio of marginal impacts on after-tax income and government costs gives the marginal value of public funds (Mayshar 1990; Hendren 2016), which we can write:

$$(8)\quad MVPF_\delta \equiv \frac{\frac{\partial B}{\partial \delta}}{\frac{\partial C}{\partial \delta}} = \frac{(1-\tau) pLATE_h}{\varphi_h - \varphi_c S_c - \tau pLATE_h}.$$

The $MVPF_\delta$ gives the value of an extra dollar spent on Head Start net of fiscal externalities. These fiscal externalities include reduced spending on competing subsidized programs (captured by the term $\varphi_c S_c$) and additional tax revenue generated by higher earnings (captured by $\tau pLATE_h$). As emphasized by Hendren (2016), the MVPF is a metric that can easily be compared across programs without specifying exactly how program expenditures are to be funded. In our case, if $MVPF_\delta > 1$, \$1 of government spending can raise the after-tax incomes of children by more than \$1, which is a robust indicator that program expansions are likely to be welfare improving.

An important lesson of the foregoing analysis is that identifying costs and benefits of changes to offer probabilities does not require identification of treatment effects relative to particular counterfactual care states. Specifically, it is not necessary to separately identify the subLATEs. This result shows that program substitution is not a design flaw of evaluations. Rather, it is a feature of the policy environment that needs to be considered when computing the likely effects of changes to policy parameters. Here, program substitution alters the usual logic of program evaluation only by requiring identification of the complier share S_c, which governs the degree of public savings realized as a result of reducing subsidies to competing programs.

V.D. Rationed Substitutes

The foregoing analysis presumes that Head Start expansions yield reductions in the enrollment of competing preschools. However, if competing programs are also oversubscribed, the slots vacated by c-compliers may be filled by other households. This will reduce the public savings associated with Head Start expansions but also generate the potential for additional test score gains.

With rationing in substitute preschool programs, the utility of enrollment in c can be written $U_i(c,Z_{ic})$, where Z_{ic} indicates an offered slot in the competing program. Household i's enrollment choice, $D_i(Z_{ih},Z_{ic})$, depends on both the Head Start offer Z_{ih} and the competing program offer. Assume these offers are assigned independently with probabilities δ_h and δ_c but that δ_c adjusts to changes in δ_h to keep total enrollment in c constant. In addition, assume that all children induced to move into c as a result of an increase in δ_c come from n rather than h.

We show in Online Appendix C that under these assumptions the marginal impact of expanding Head Start becomes:

$$\frac{\partial E[Y_i]}{\partial \delta_h} = (LATE_h + LATE_{nc} \cdot S_c) \times \frac{\partial P(D_i = h)}{\partial \delta_h},$$

where $LATE_{nc} \equiv E[Y_i(c) - Y_i(n) | D_i(Z_{ih},1) = c, D_i(Z_{ih},0) = n]$. Intuitively, every c-complier now spawns a corresponding n-to-c complier who fills the vacated preschool slot.

The marginal cost to government of inducing this change in test scores can be written:

$$\frac{\partial C}{\partial \delta_h} = \left[\varphi_h - \tau p(LATE_h + LATE_{nc} \cdot S_c)\right] \times \frac{\partial P(D_i = h)}{\partial \delta_h}.$$

Relative to equation (7), rationing eliminates the public savings from reduced enrollment in substitute programs but adds another fiscal externality in its place: the tax revenue associated with any test score gains of shifting children from home care to competing preschools. The resulting marginal value of public funds can be written:

$$MVPF_{\delta,rat} = \frac{(1-\tau)p(LATE_h + LATE_{nc} \cdot S_c)}{\varphi_h - \tau p(LATE_h + LATE_{nc} \cdot S_c)}. \tag{9}$$

While the impact of rationed substitutes on the marginal value of public funds is theoretically ambiguous, there is good reason to expect $MVPF_{\delta,rat} > MVPF_{\delta}$ in practice. Specifically, ignoring rationing of competing programs yields a lower bound on the rate of return to Head Start expansions if Head Start and other forms of center-based care have roughly comparable effects on test scores and competing programs are cheaper (see Online Appendix C). Unfortunately, effects for n-to-c compliers are not nonparametrically identified by the HSIS experiment because one cannot know which households that care for their children at home would otherwise choose to enroll them in competing preschools. We return to this issue in Section IX.

V.E. Structural Reforms

An important assumption in the previous analyses is that changing lottery probabilities does not alter the mix of program compliers. Consider now the effects of altering some structural feature f of the Head Start program that households value but has no impact on test scores. For example, Executive Order 13330, issued by President George W. Bush in February 2004, mandated enhancements to the transportation services provided by Head Start and other federal programs (Federal Register 2004). Expanding Head Start transportation services should not directly influence educational outcomes but may yield a compositional effect by drawing in households from a different mix of counterfactual care environments.[10] By shifting the composition

10. This presumes that peer effects are not an important determinant of test score outcomes. Large changes in the student composition of Head Start classrooms could potentially change the effectiveness of Head Start.

of program participants, changes in f may boost the program's rate of return.

To establish notation, we assume that households now value Head Start participation as:

$$\tilde{U}_i(h, Z_i, f) = U_i(h, Z_i) + f.$$

Utilities for other preschools and home care are assumed to be unaffected by changes in f. This implies that increases in f make Head Start more attractive for all households. For simplicity, we return to our prior assumption that competing programs are not rationed. As shown in Online Appendix C, the assumption that f has no effect on potential outcomes implies:

$$\frac{\partial E[Y_i]}{\partial f} = MTE_h \times \frac{\partial P(D_i = h)}{\partial f},$$

where

$$MTE_h \equiv E[Y_i(h) - Y_i(c) | U_i(h, Z_i) + f = U_i(c), U_i(c) > U_i(n)]\vec{S}_c$$
$$+ E[Y_i(h) - Y_i(n) | U_i(h, Z_i) + f = U_i(n), U_i(n) > U_i(c)](1 - \vec{S}_c),$$

and $\vec{S}_c$ gives the share of children on the margin of participating in Head Start who prefer the competing program to preschool nonparticipation. Following the terminology in Heckman, Urzua, and Vytlacil (2008), the marginal treatment effect MTE_h is the average effect of Head Start on test scores among households indifferent between Head Start and the next best alternative. This is a marginal version of the result in equation (6), where integration is now over a set of children who may differ from current program compliers in their mean impacts. Like $LATE_h$, MTE_h is a weighted average of "subMTEs" corresponding to whether the next best alternative is home care or a competing preschool program. The weight $\vec{S}_c$ may differ from S_c if inframarginal participants are drawn from different sources than marginal ones.

The test score effects of improvements to the program feature must be balanced against the costs. We suppose that changing program features changes the *average* cost $\varphi_h(f)$ of Head Start services, so that the net costs to government of financing preschool are now:

$$C(f) = C_0 + \varphi_h(f)P(D_i = h) + \varphi_c P(D_i = c) - \tau pE[Y_i], \tag{10}$$

where $\frac{\partial \varphi_h(f)}{\partial f} \geq 0$. The marginal costs to government (per program complier) of a change in the program feature can be written:

$$\frac{\frac{\partial C(f)}{\partial f}}{\frac{\partial P(D_i = h)}{\partial f}} = \underbrace{\varphi_h}_{\text{Marginal Provision Cost}} + \underbrace{\frac{\frac{\partial \varphi_h(f)}{\partial f}}{\frac{\partial ln\, P(D_i = h)}{\partial f}}}_{\text{Inframarginal Provision Cost}} - \underbrace{\varphi_c \vec{S}_c}_{\text{Public Savings}} - \underbrace{\tau pMTE_h}_{\text{Added Revenue}} \tag{11}$$

The first term on the right-hand side of equation (11) gives the administrative cost of enrolling another child. The second term gives the increased cost of providing inframarginal families with the improved program feature. The third term is the expected savings in reduced funding to competing preschool programs. The final term gives the additional tax revenue raised by the boost in the marginal enrollee's human capital.

Letting $\eta \equiv \frac{\frac{\partial ln\, \varphi(f)}{\partial f}}{\frac{\partial ln\, P(D_i = h)}{\partial f}}$ be the elasticity of costs with respect to enrollment, we can write the marginal value of public funds associated with a change in program features as:

$$MVPF_f \equiv \frac{\frac{\partial B}{\partial f}}{\frac{\partial C(f)}{\partial f}} = \frac{(1-\tau)pMTE_h}{\varphi_h(1+\eta) - \varphi_c \vec{S}_c - \tau pMTE_h}. \tag{12}$$

As in our analysis of optimal program scale, equation (11) shows that it is not necessary to separately identify the subMTEs that compose MTE_h to determine the optimal value of f. Rather, it is sufficient to identify the average causal effect of Head Start for children on the margin of participation along with the average net cost of an additional seat in this population.

VI. A Cost-Benefit Analysis of Program Expansion

We use the HSIS data to conduct a formal cost-benefit analysis of changes to Head Start's offer rate under the assumption that competing programs are not rationed (we consider the case with rationing in Section IX). Our analysis focuses on the costs

and benefits associated with one year of Head Start attendance.[11] This exercise requires estimates of each term in equation (7). We estimate $LATE_h$ and S_c from the HSIS and calibrate the remaining parameters using estimates from the literature. Calibrated parameters are listed in Table IV, Panel A. To be conservative, we deliberately bias our calibrations toward understating Head Start's benefits and overstating its costs to arrive at a lower bound rate of return. Further details of the calibration exercise are provided in Online Appendix D.

Table IV, Panel B reports estimates of the marginal value of public funds associated with an expansion of Head Start offers ($MVPF_\delta$). To account for sampling uncertainty in our estimates of $LATE_h$ and S_c, we report standard errors calculated via the delta method. Because asymptotic delta method approximations can be inaccurate when the statistic of interest is highly nonlinear (Lafontaine and White 1986), we also report bootstrap p-values from one-tailed tests of the null hypothesis that the benefit-cost ratio is less than 1.[12]

The results show that accounting for the public savings associated with enrollment in substitute preschools has a large effect on the estimated social value of Head Start. We conduct cost-benefit analyses under three assumptions: φ_c is either 0, 50%, or 75% of φ_h. Our preferred calibration uses $\varphi_c = 0.75\varphi_h$, reflecting that fact that roughly 75% of competing centers are publicly funded (see Online Appendix D). Setting $\varphi_c = 0$ yields a $MVPF_\delta$ of 1.10. Setting φ_c equal to $0.5\varphi_h$ and $0.75\varphi_h$ raises the $MVPF_\delta$ to 1.50 and 1.84, respectively. This indicates that the fiscal externality generated by program substitution has an important effect on the social value of Head Start. Bootstrap tests decisively reject values of $MVPF_\delta$ less than 1 when $\varphi_c = 0.5\varphi_h$ or $0.75\varphi_h$.

11. Children in the three-year-old cohort who enroll for two years generate additional costs. As shown in Table III, a Head Start offer raises the probability of enrollment in the second year by only 0.16, implying that first-year offers have modest net effects on second-year costs. Enrollment for two years may also generate additional benefits, but these cannot be estimated without strong assumptions on the Head Start dose-response function. We therefore consider only first-year benefits and costs.

12. This test is computed by a nonparametric block bootstrap of the Studentized t-statistic that resamples Head Start sites. We have found in Monte Carlo exercises that delta method confidence intervals for $MVPF_\delta$ tend to overcover, whereas bootstrap-t tests have approximately correct size. This is in accord with theoretical results from Hall (1992) that show bootstrap-t methods yield a higher-order refinement to p-values based on the standard delta method approximation.

TABLE IV

BENEFITS AND COSTS OF HEAD START

(1) Parameter	(2) Description	(3) Value	(4) Source
Panel A: Parameter values			
p	Effect of a 1 std. dev. increase in test scores on earnings	$0.1\bar{e}$	Table A.IV
e_{US}	U.S. average present discounted value of lifetime earnings at age 3.4	\$438,000	Chetty et al. (2011) with 3% discount rate
e_{parent}/e_{US}	Average earnings of Head Start parents relative to U.S. average	0.46	Head Start Program Facts
IGE	Intergenerational income elasticity	0.40	Lee and Solon (2009)
$\bar{e}$	Average present discounted value of lifetime earnings for Head Start applicants	\$343,392	$[1-(1-e_{parent}/e_{US})IGE]e_{US}$
$0.1\bar{e}$	Effect of a 1 std. dev. increase in test scores on earnings of Head Start applicants	\$34,339	
$LATE_h$	Local average treatment effect	0.247	HSIS
τ	Marginal tax rate for Head Start population	0.35	CBO (2012)
S_c	Share of Head Start population drawn from other preschools	0.34	HSIS
φ_h	Marginal cost of enrollment in Head Start	\$8,000	Head Start Program Facts
φ_c	Marginal cost of enrollment in other preschools	\$0	Naive assumption: $\varphi_c = 0$
		\$4,000	Pessimistic assumption: $\varphi_c = 0.5\varphi_h$
		\$6,000	Preferred assumption: $\varphi_c = 0.75\varphi_h$

TABLE IV
(CONTINUED)

(1) Parameter	(2) Description	(3) Value	(4) Source
Panel B: Marginal value of public funds			
NMB	Marginal benefit to Head Start population net of taxes	\$5,513	$(1 - \tau)pLATE_h$
MFC	Marginal fiscal cost of Head Start enrollment	\$5,031	$\varphi_h - \varphi_c S_c - \tau pLATE_h$, naive assumption
		\$3,671	Pessimistic assumption
		\$2,991	Preferred assumption
MVPF	Marginal value of public funds	1.10 (0.22)	*NMB*/*MFC* (std. err.), naive assumption
		p-value = .1	
		Breakeven $p/\bar{e} = 0.09\,(0.01)$	
		1.50 (0.34)	Pessimistic assumption
		p-value = .00	
		Breakeven $p/\bar{e} = 0.08\,(0.01)$	
		1.84 (0.47)	Preferred assumption
		p-value = .00	
		Breakeven $p/\bar{e} = 0.07\,(0.01)$	

Notes. This table reports results of cost-benefit calculations for Head Start. Parameter values are obtained from the sources listed in column (4). Standard errors for *MVPF* ratios are calculated using the delta method. *P*-values are from one-tailed tests of the null hypotheses that the *MVPF* is less than 1. These tests are performed via nonparametric block bootstrap of the t-statistic, clustered at the Head Start center level. Breakevens give percentage effects of a standard deviation of test scores on earnings that set *MVPF* equal to 1.

Notably, our preferred estimate of 1.84 is well above the estimated rates of return of comparable expenditure programs summarized in Hendren (2016), and comparable to the marginal value of public funds associated with increases in the top marginal tax rate (between 1.33 and 2.0).

To assess the sensitivity of our results to alternative assumptions regarding the relationship between test score effects and earnings, Table IV also reports breakeven relationships between test scores and earnings that set $MVPF_\delta$ equal to 1 for each value of φ_c. When $\varphi_c = 0$ the breakeven earnings effect is 9% per test score standard deviation, only slightly below our calibrated value of 10%. This indicates that when substitution is ignored, Head Start is close to breaking even, and small changes in assumptions will yield values of $MVPF_\delta$ below 1. Increasing φ_c to $0.5\varphi_h$ or $0.75\varphi_h$ reduces the breakeven earnings effect to 8% or 7%, respectively. The latter figure is well below comparable estimates in the recent literature, such as estimates from Chetty et al.'s (2011) study of the Tennessee STAR class size experiment (13%; see Appendix Table A.IV). Therefore, after accounting for fiscal externalities, Head Start's costs are estimated to exceed its benefits only if its test score impacts translate into earnings gains at a lower rate than similar interventions for which earnings data are available.

VII. Beyond LATE

Thus far, we have evaluated the return to a marginal expansion of Head Start under the assumption that the mix of compliers can be held constant. However, it is likely that major reforms to Head Start would entail changes to program features such as accessibility that could in turn change the mix of program compliers. To evaluate such reforms, it is necessary to predict how selection into Head Start is likely to change and how this affects the program's rate of return.

VII.A. IV Estimates of SubLATEs

A first way in which selection into Head Start could change is if the mix of compliers drawn from home care and competing preschools were altered while holding the composition of those two groups constant. To predict the effects of such a change on the program's rate of return we need to estimate the subLATEs in equation (3).

One approach to identifying subLATEs is to conduct two-stage least squares (2SLS) estimation treating Head Start enrollment and enrollment in other preschools as separate endogenous variables. A common strategy for generating instruments in such settings is to interact an experimentally assigned program offer with observed covariates or site indicators (e.g., Kling, Liebman, and Katz 2007; Abdulkadiroğlu, Angrist, and Pathak 2014). Such approaches can secure identification in a constant effects framework but, as we demonstrate in Online Appendix E, will typically fail to identify interpretable parameters if the subLATEs themselves vary across the interacting groups (see Hull 2015 and Kirkeboen, Leuven, and Mogstad 2016 for related results).

Table V reports 2SLS estimates of the separate effects of Head Start and competing preschools using as instruments the Head Start offer indicator and its interaction with eight student- and site-level covariates likely to capture heterogeneity in compliance patterns.[13] These instruments strongly predict Head Start enrollment but induce relatively weak independent variation in enrollment in other preschools, with a partial first-stage F-statistic of only 1.8. The 2SLS estimates indicate that Head Start and other centers yield large and roughly equivalent effects on test scores of approximately 0.4 standard deviations. This finding is roughly in line with the view that preschool effects are homogeneous and that program substitution simply attenuates IV estimates of the effect of Head Start relative to home care. Cautioning against this interpretation is the 2SLS overidentification test, which strongly rejects the constant effects model, indicating the presence of substantial effect heterogeneity across covariate groups.

A separate source of variation comes from experimental sites: the HSIS was implemented at hundreds of individual Head Start centers, and previous studies have shown substantial variation in treatment effects across these centers (Bloom and Weiland 2015;

13. Previous analyses of the HSIS have shown important effect heterogeneity with respect to baseline scores and first language (Bitler, Domina, and Hoynes 2014; Bloom and Weiland 2015), so we include these in the list of student-level interactions. We also allow interactions with variables measuring whether a child's center of random assignment offers transportation to Head Start, whether the center of random assignment is above the median of the Head Start quality measure, the education level of the child's mother, whether the child is age four, whether the child is black, and an indicator for family income above the poverty line.

TABLE V

TWO-STAGE LEAST SQUARES ESTIMATES WITH INTERACTION INSTRUMENTS

	One endogenous variable:	Two endogenous variables	
Instruments	Head Start (1)	Head Start (2)	Other centers (3)
Offer	0.247		
(1 instrument)	(0.031)		
Offer × covariates	0.241	0.384	0.419
(9 instruments)	(0.030)	(0.127)	(0.359)
First-stage *F*	276.2	17.7	1.8
Overid. *p*-value	.007	.006	
Offer × sites	0.210	0.213	0.008
(183 instruments)	(0.026)	(0.039)	(0.095)
First-stage *F*	215.1	90.0	2.7
Overid. *p*-value	.002	.002	
Offer × site groups	0.229	0.265	0.110
(6 instruments)	(0.029)	(0.056)	(0.146)
First-stage *F*	1,015.2	339.1	32.6
Overid. *p*-value	.077	.050	
Offer × covariates and offer × site groups (14 instruments)	0.229 (0.029)	0.302 (0.054)	0.225 (0.134)
First-stage *F*	340.2	121.2	13.3
Overid. *p*-value	.012	.010	

Notes. This table reports two-stage least squares estimates of the effects of Head Start and other preschool centers in spring 2003. The model in the first row instruments Head Start attendance with the Head Start offer. Models in the second row instrument Head Start and other preschool attendance with interactions of the offer and transportation, above-median quality, race, Spanish language, mother's education, an indicator for income above the federal poverty line, and baseline score. The third row uses the Head Start offer interacted with 183 experimental site indicators as instruments. The fourth row uses interactions of the offer and indicators for groups of experimental sites obtained from a multinomial probit model with unobserved group fixed effects, as described in Online Appendix G. The fifth row uses both covariate and site group interactions. All models control for main effects of the interacting variables and baseline covariates. First-stage *F*-statistics are Angrist and Pischke (2009) partial *F*'s. Standard errors are clustered at the center level.

Walters 2015). Using site interactions as instruments again yields much more independent variation in Head Start enrollment than in competing preschools.[14] However, the estimated impact of Head Start is smaller, and competing centers are estimated to yield no gains relative to home care. While these site-based estimates are nominally more precise than those obtained from the covariate interactions, with 183 instruments the

14. To avoid extreme imbalance in site size, we grouped the 356 sites in our data into 183 sites with 10 or more observations. See Online Appendix G for details.

asymptotic standard errors may provide a poor guide to the degree of uncertainty in the parameter estimates (Bound, Jaeger, and Baker 1995). We explore this issue in Online Appendix Table A.V, which reports limited information maximum likelihood and jackknife IV estimates of the same model. These approaches yield much larger standard errors and very different point estimates, suggesting that weak instrument biases are at play here.

To deal with these statistical problems, we use a choice model with discrete unobserved heterogeneity (described in more detail later) to aggregate Head Start sites together into six groups with similar substitution patterns. Using the site group interactions as instruments yields significant independent variation in both Head Start and competing preschool enrollment and produces results more in line with those obtained from the covariate interactions. Pooling the site group and covariate interaction instruments together yields the most precise estimates, which indicate that both preschool types increase scores relative to home care and that Head Start is slightly more effective than competing preschools. However, the overidentification test continues to reject the constant effects model, suggesting that these estimates are still likely to provide a misleading guide to the underlying subLATEs. Another important limitation of the interacted 2SLS approach is that it conditions on realized selection patterns and therefore cannot be used to predict the effects of reforms that change the underlying composition of n- and c-compliers. We now turn to developing an econometric selection model that allows us to address both of these limitations.

VII.B. Selection Model

Our selection model parameterizes the preferences and potential outcomes introduced in the model of Section V to motivate a two-step control function estimator. Like the interacted 2SLS approach, the proposed estimator exploits interactions of the Head Start offer with covariates and site groups to separately identify the causal effects of care alternatives. Unlike the interacted 2SLS approach, the control function estimator allows the interacting groups to have different subLATEs that vary parametrically with the probability of enrolling in Head Start and competing preschools.

Normalizing the value of preschool nonparticipation to zero, we assume households have utilities over program alternatives given by:

$$\begin{aligned} U_i(h, Z_i) &= \psi_h(X_i, Z_i) + v_{ih}, \\ U_i(c) &= \psi_c(X_i) + v_{ic}, \\ U_i(n) &= 0, \end{aligned} \tag{13}$$

where X_i denotes the vector of baseline household and experimental site characteristics listed in Table I and Z_i again denotes the Head Start offer dummy. The stochastic components of utility (v_{ih}, v_{ic}) reflect unobserved differences in household demand for Head Start and competing preschools relative to home care. In addition to pure preference heterogeneity, these terms may capture unobserved constraints such as whether family members are available to help with child care. We suppose these components obey a multinomial probit specification:

$$(v_{ih}, v_{ic})|X_i, Z_i \sim N\left(0, \begin{bmatrix} 1 & \rho(X_i) \\ \rho(X_i) & 1 \end{bmatrix}\right),$$

which allows for violations of the independence from irrelevant alternatives (IIA) condition that underlies multinomial logit selection models such as that of Dubin and McFadden (1984).

As in the Heckman (1979) selection framework, we model endogeneity in participation decisions by allowing linear dependence of mean potential outcomes on the unobservables that influence choices. Specifically, for each program alternative $d \in \{h,c,n\}$, we assume:

$$E[Y_i(d)|X_i, Z_i, v_{ih}, v_{ic}] = \mu_d(X_i) + \gamma_{dh} v_{ih} + \gamma_{dc} v_{ic}. \tag{14}$$

The $\{\gamma_{dh}, \gamma_{dc}\}$ coefficients in equation (14) describe the nature of selection on unobservables. This specification can accommodate a variety of selection schemes. For example, if $\gamma_{dh} = \gamma_h > 0$, then conditional on observables, selection into Head Start is governed by potential outcome levels—those most likely to participate in Head Start have higher test scores in all care environments. But if $\gamma_{hh} > 0$ and $\gamma_{nh} = -\gamma_{hh}$, then households engage in Roy (1951)-style selection into Head Start based on test score gains—those most likely to participate in Head Start receive larger test score benefits when they switch from home care to Head Start.

By iterated expectations, equation (14) implies the conditional expectation of realized outcomes can be written:

$$E[Y_i|X_i,Z_i,D_i = d] = \mu_d(X_i) + \gamma_{dh}\lambda_h(X_i,Z_i,d) + \gamma_{dc}\lambda_c(X_i,Z_i,d), \tag{15}$$

where $\lambda_h(X_i,Z_i,D_i) \equiv E[v_{ih}|X_i,Z_i,D_i]$ and $\lambda_c(X_i, Z_i, D_i) \equiv E[v_{ic}|X_i,Z_i,D_i]$ are generalizations of the standard inverse Mills ratio terms used in the two-step Heckman (1979) selection correction (see Online Appendix F for details). These terms depend on X_i and Z_i only through the conditional probabilities of enrolling in Head Start and other preschools.

VII.C. Identification

To demonstrate identification of the selection coefficients $\{\gamma_{dh}, \gamma_{dc}\}$ it is useful to eliminate the main effect of the covariates by differencing equation (15) across values of the program offer Z_i as follows:

$$\begin{aligned} &E[Y_i|X_i,Z_i = 1,D_i = d] - E[Y_i|X_i,Z_i = 0,D_i = d] \\ &= \gamma_{dh}[\lambda_h(X_i,1,d) - \lambda_h(X_i,0,d)] + \gamma_{dc}[\lambda_c(X_i,1,d) - \lambda_c(X_i,0,d)]. \end{aligned} \tag{16}$$

This difference measures how *selected* test score outcomes in a particular care alternative respond to a Head Start offer. Responses in selected outcomes are driven entirely by compositional changes—that is, from compliers switching between alternatives.

With two values of the covariates X_i, equation (16) can be evaluated twice, yielding two equations in the two unknown selection coefficients. Online Appendix F details the conditions under which this system can be solved and provides expressions for the selection coefficients in terms of population moments. Additive separability of the potential outcomes in observables and unobservables is essential for identification. If the selection coefficients in equation (16) were allowed to depend on X_i, there would be two unknowns for every value of the covariates and identification would fail. Heuristically, then, our key assumption is that selection on unobservables works "the same way" for every value of the covariates, which allows us to exploit variation in selected outcome responses across subgroups to infer the parameters governing the selection process.

To understand this restriction, suppose (as turns out to be the case) that college-educated mothers are more likely to enroll their children in competing preschools when denied access to Head Start. Our model allows Head Start and other preschools to have different average treatment effects on the children of more and less educated mothers. However, it rules out the possibility that children with college-educated mothers sort into Head Start on the basis of potential test score gains, while children of less educated mothers exhibit no sorting on these gains. As in Brinch, Mogstad, and Wiswall (forthcoming), this restriction is testable when X_i takes more than two values because it implies we should obtain similar estimates of the selection coefficients based on variation in different subsets of the covariates. We provide evidence along these lines by contrasting estimates that exploit site variation with estimates based on household covariates.

VII.D. Estimation

To make estimation tractable, we approximate $\psi_h(X, Z)$ and $\psi_c(X)$ with flexible linear functions. The nonseparability of $\psi_h(X, Z)$ is captured by linear interactions between Z and the eight covariates used in our earlier 2SLS analysis. We also allow interactions with the 183 experimental site indicators, but to avoid incidental parameters problems, we constrain the coefficients on those dummies to belong to one of K discrete categories. Results from Bonhomme and Manresa (2015) and Saggio (2012) suggest that this "grouped fixed effects" approach should yield good finite sample performance even when some sites have as few as 10 observations. As described in Online Appendix G, we choose the number of site groups K using the Bayesian information criterion (BIC). Finally, all of the interacting variables (both site groups and covariates) are allowed to influence the correlation parameter $\rho(X)$. We assume that $\operatorname{arctanh}\rho(X) = \frac{1}{2}\ln\left(\frac{1+\rho(X)}{1-\rho(X)}\right)$ is linear in these variables, a standard transformation that ensures the correlation is between -1 and 1 (Cox 2008).

The model is fit in two steps. First, we estimate the parameters of the probit model via simulated maximum likelihood, evaluating choice probabilities with the Geweke-Hajivassiliou-Keane (GHK) simulator (Geweke 1989; Keane 1994; Hajivassiliou and McFadden 1998). Models including site groups are estimated with an algorithm that alternates between maximizing the likelihood and reassigning groups, described in detail in Online

Appendix G. Second, we use the parameters of the choice model to form control function estimates $\left(\hat{\lambda}_h(X_i,Z_i,D_i), \hat{\lambda}_c(X_i,Z_i,D_i)\right)$, which are then used in a second step regression of the form:

$$
\begin{aligned}
Y_i = {} & \theta_{n0} + X'\theta_{nx} + \gamma_{nh}\hat{\lambda}_h(X_i,Z_i,D_i) + \gamma_{nc}\hat{\lambda}_c(X_i,Z_i,D_i) \\
& + 1\{D_i=c\}\begin{bmatrix}(\theta_{c0}-\theta_{n0}) + X_i'(\theta_{cx}-\theta_{nx}) + (\gamma_{ch}-\gamma_{nh})\hat{\lambda}_h(X_i,Z_i,c) \\ + (\gamma_{cc}-\gamma_{nc})\hat{\lambda}_c(X_i,Z_i,c)\end{bmatrix} \qquad (17) \\
& + 1\{D_i=h\}\begin{bmatrix}(\theta_{h0}-\theta_{n0}) + X_i'(\theta_{hx}-\theta_{nx}) + (\gamma_{hh}-\gamma_{nh})\hat{\lambda}_h(X_i,Z_i,h) \\ + (\gamma_{hc}-\gamma_{nc})\hat{\lambda}_c(X_i,Z_i,h)\end{bmatrix} + \epsilon_i.
\end{aligned}
$$

The covariate vector X_i is normed to have unconditional mean zero, so the intercepts θ_{d0} can be interpreted as average potential outcomes. Hence, the differences $\theta_{h0} - \theta_{n0}$ and $\theta_{h0} - \theta_{c0}$ capture average treatment effects of Head Start and other preschools relative to no preschool. To avoid overfitting, we restrict variables other than the site types and eight key covariates to have common coefficients across care alternatives.[15] Inference on the second-step parameters is conducted via the nonparametric block bootstrap, clustered by experimental site.

VIII. Model Estimates

VIII.A. Model Parameters

Table VI reports estimates of the full choice model obtained from exploiting both covariates and site heterogeneity. The BIC selects a specification with six site groups for the full model (see Online Appendix Table A.VI), with group shares that vary between 12% and 21% of the sample. These assignments make up the site groups used in the earlier 2SLS analysis of Table V.

Columns (1) and (2) of Table VI show the coefficients governing the mean utility of enrollment in Head Start. We easily reject the null hypothesis that the program offer interaction effects in the Head Start utility equation are homogeneous. Panel A, column (2) indicates that the effects of an offer are greater at high-quality centers and lower among nonpoor children that would

15. This restriction cannot be statistically rejected and has minimal effects on the point estimates.

TABLE VI

MULTINOMIAL PROBIT ESTIMATES

	Head Start utility			
	(1) Main effect	(2) Offer interaction	(3) Other center utility	(4) Arctanh ρ
Panel A: Covariates				
Center provides	0.022	0.111	0.054	0.096
transportation	(0.114)	(0.142)	(0.087)	(0.178)
Above-median	−0.233	0.425	−0.115	−0.007
center quality	(0.091)	(0.102)	(0.082)	(0.153)
Black	0.095	0.282	0.206	−0.185
	(0.108)	(0.127)	(0.100)	(0.166)
Spanish speaker	−0.049	−0.273	−0.213	0.262
	(0.136)	(0.122)	(0.169)	(0.224)
Mother's education	0.106	0.021	0.105	−0.219
	(0.056)	(0.060)	(0.064)	(0.110)
Income above FPL	0.216	−0.305	0.173	0.097
	(0.128)	(0.121)	(0.126)	(0.192)
Baseline score	0.080	−0.025	0.292	0.026
	(0.094)	(0.108)	(0.069)	(0.094)
Age 4	0.164	−0.277	0.518	0.010
	(0.142)	(0.166)	(0.104)	(0.170)
p-values: no heterogeneity	.015	.000	.000	.666
Panel B: Experimental site groups				
Group 1	−0.644	2.095	0.424	0.435
(share = 0.215)	(0.136)	(0.153)	(0.085)	(0.128)
Group 2	−4.847	6.760	−0.577	−0.496
(share = 0.183)	(0.076)	(0.158)	(0.045)	(0.172)
Group 3	−2.148	2.912	−0.768	0.530
(share = 0.183)	(0.312)	(0.340)	(0.081)	(0.159)
Group 4	0.488	0.541	−0.139	0.483
(share = 0.151)	(0.130)	(0.150)	(0.226)	(0.322)
Group 5	−1.243	2.849	−1.643	−0.772
(share = 0.145)	(0.108)	(0.171)	(0.164)	(0.359)
Group 6	0.072	1.191	0.110	2.988
(share = 0.124)	(0.127)	(0.183)	(0.106)	(0.925)
p-values: no heterogeneity	.000	.000	.000	.000

Notes. This table reports simulated maximum likelihood estimates of a multinomial probit model of preschool choice. The model includes fixed effects for six unobserved groups of experimental sites, estimated as described in Online Appendix G. The Head Start and other center utilities also include the main effects of sex, test language, teen mother, mother's marital status, presence of both parents, family size, special education, family income categories, and second- and third-order terms in baseline test scores. The likelihood is evaluated using the GHK simulator, and likelihood contributions are weighted by the reciprocal of the probability of experimental assignments. *P*-values for site heterogeneity are from tests that all group-specific constants are equal. *P*-values for covariate heterogeneity are from tests that all covariate coefficients in a column are 0. Standard errors are clustered at the Head Start center level.

typically be ineligible for Head Start enrollment.[16] Panel B, column (2) reveals the presence of significant heterogeneity across site groups in the response to a program offer, which likely reflects unobserved market features such as the presence or absence of state provided preschool.

Column (4) of Table VI reports the parameters governing the correlation in unobserved tastes for Head Start and competing programs. The correlation is positive for four of six site groups, indicating that most households view preschool alternatives as more similar to each other than to home care. This establishes that the IIA condition underlying logit-based choice models is empirically violated. Although there is some evidence of heterogeneity in the correlation based on mother's education, we cannot reject the joint null hypothesis that the correlation is constant across covariate groups. However, we easily reject that the correlation is constant across site groups.

The many sources of heterogeneity captured by the choice model yield substantial variation in predicted enrollment shares for Head Start and competing preschools. Online Appendix Figure A.I shows that these predictions match variation in choice probabilities across subgroups. Moreover, diagnostics indicate this variation is adequate to secure separate identification of the second stage control function coefficients. From equation (16), the model is underidentified if, for any alternative d, the control function difference $\lambda_h(X_i,1,d) - \lambda_h(X_i,0,d)$ is linearly dependent on the corresponding difference $\lambda_c(X_i,1,d) - \lambda_c(X_i,0,d)$. Online Appendix Figure A.II shows that the deviations from linear dependence are visually apparent and strongly statistically significant.

Table VII reports second-step estimates of the parameters in equation (17). Column (1) omits all controls and simply reports differences in mean test scores across care alternatives (the omitted category is home care). Head Start students score 0.2 standard deviations higher than students in home care, and the corresponding difference for students in competing preschools is 0.26 standard deviations. Column (2) adds controls for baseline

16. The quality variable aggregates information on center characteristics (teacher and center director education and qualifications, class size) and practices (variety of literacy and math activities, home visiting, health and nutrition) measured in interviews with center directors, teachers, and parents of children enrolled in the preschool center.

TABLE VII

SELECTION-CORRECTED ESTIMATES OF PRESCHOOL EFFECTS

	Least squares		Control function		
	(1) No controls	(2) Covariates	(3) Covariates	(4) Site groups	(5) Full model
Head Start	0.202	0.218	0.483	0.380	0.470
	(0.037)	(0.022)	(0.117)	(0.121)	(0.101)
Other preschools	0.262	0.151	0.183	0.065	0.109
	(0.052)	(0.035)	(0.269)	(0.991)	(0.253)
λ_h			0.015	0.004	0.019
			(0.053)	(0.063)	(0.053)
Head Start $\times$ λ_h			−0.167	−0.137	−0.158
			(0.080)	(0.126)	(0.091)
Other preschools $\times$ λ_h			−0.030	−0.047	0.000
			(0.109)	(0.366)	(0.115)
λ_c			−0.333	−0.174	−0.293
			(0.203)	(0.187)	(0.115)
Head Start $\times$ λ_c			0.224	0.065	0.131
			(0.306)	(0.453)	(0.172)
Other preschools $\times$ λ_c			0.488	0.440	0.486
			(0.248)	(0.926)	(0.197)
p-values:					
No selection			.016	.510	.046
No selection on gains			.133	.560	.084
Additive separability			.261	.452	.349

Notes. This table reports selection-corrected estimates of the effects of Head Start and other preschool centers in spring 2003. Each column shows coefficients from regressions of test scores on an intercept, a Head Start indicator, an other preschool indicator, and controls. Column (1) shows estimates with no controls. Column (2) adds controls for sex, race, home language, test language, mother's education, teen mother, mother's marital status, presence of both parents, family size, special education, income categories, experimental site characteristics (transportation, above-median quality, and urban status) and a third-order polynomial in baseline test score. This column interacts the preschool variables with transportation, above-median quality, race, Spanish language, mother's education, an indicator for income above the federal poverty line, and the main effect of baseline score. Covariates are demeaned in the estimation sample, so that main effects can be interpreted as estimates of average treatment effects. Column (3) adds control function terms constructed from a multinomial probit model using the covariates from column (2) and the Head Start offer. The interacting variables from column (2) are allowed to interact with the Head Start offer and enter the preschool taste correlation equation in column (3). Column (4) omits observed covariates and includes indicators for experimental site groups, constructed using the algorithm described in Online Appendix G. The multinomial probit model is saturated in these site group indicators, and the second-step regression interacts site groups with preschool alternatives. Column (5) combines the variables used in columns (3) and (4). Standard errors are bootstrapped and clustered at the center level. The bottom row shows p-values from a score test of the hypothesis that interactions between the control functions and covariates are 0 in each preschool alternative (see Online Appendix F for details).

characteristics. Because the controls include a third-order polynomial in baseline test scores, this column can be thought of as reporting "value-added" estimates of the sort that have received renewed attention in the education literature (Kane, Rockoff, and Staiger 2008; Rothstein 2010; Chetty, Friedman, and Rockoff 2014a). Surprisingly, adding these controls does little to the estimated effect of Head Start relative to home care but improves precision. By contrast, the estimated impact of competing preschools relative to home care falls significantly once controls are added.

Columns (3)–(5) add control functions adjusting for selection on unobservables based on choice models with covariates, site groups, or both. Unlike the specifications in previous columns, these control function terms exploit experimental variation in offer assignment. Adjusting for selection on unobservables dramatically raises the estimated average impact of Head Start relative to home care. However, the estimates are fairly imprecise. Imprecision in estimates of average treatment effects is to be expected given that these quantities are only identified via parametric restrictions that allow us to infer the counterfactual outcomes of always takers and never takers. Below we consider average treatment effects on compliers, which are estimated more precisely.

Although some of the control function coefficient estimates are also imprecise, we reject the hypotheses of no selection on levels ($\gamma_{kd} = 0\ \forall(k,d)$) and no selection on gains ($\gamma_{dk} = \gamma_{jk}$ for $d \neq j$, $k \in \{h,c\}$) in our most precise specification. The selection coefficient estimates exhibit some interesting patterns. One regularity is that estimates of $\gamma_{hh} - \gamma_{nh}$ are negative in all specifications (though insignificant in the model using site groups only). In other words, children who are more likely to attend Head Start receive smaller achievement benefits when shifted from home care to Head Start. This "reverse Roy" pattern of negative selection on test score gains suggests large benefits for children with unobservables making them less likely to attend the program.[17] Other preschool programs, by contrast, seem to exhibit positive selection on gains: the estimated difference $\gamma_{cc} - \gamma_{nc}$ is always

17. Walters (2014) finds a related pattern of negative selection in the context of charter schools, though in his setting the fall-back potential outcome (as opposed to the charter school outcome) appears to respond positively to unobserved characteristics driving program participation.

positive and is significant in the full model. A possible interpretation of these patterns is that Head Start is viewed by parents as a preschool of last resort, leading to enrollment by the families most desperate to get help with child care. Such households cannot be selective about whether the local Head Start center is a good match for their child, which results in lower test score gains. By contrast, households considering enrollment in substitute preschools may have greater resources that afford them the luxury of being more selective about whether such programs are a good match for their child.

Estimates of the control function coefficients are very similar in columns (3) and (4), though the estimates are less precise when only site group interactions are used. This indicates that the implied nature of selection is the same regardless of whether identification is based on site or covariate interactions, lending credibility to our assumption that selection works the same way across subgroups. Also supporting this assumption are the results of score tests for the additive separability of control functions and covariates, reported in the bottom row of Table VII. These tests are conducted by regressing residuals from the two-step models on interactions of the control functions with covariates and site groups, along with the main effects from equation (15). In all specifications, we fail to reject additive separability at conventional levels (see Online Appendix F for some additional goodness-of-fit tests). Although these tests do not have the power to detect all forms of nonseparability, the correspondence between estimates based on covariate and site variation suggests that our key identifying assumption is reasonable.

VIII.B. Treatment Effects

Table VIII reports average treatment effects on compliers for each of our selection-corrected models. The first row uses the model parameters to compute the pooled $LATE_h$, which is nonparametrically identified by the experiment. The model estimates line up closely with the nonparametric estimate obtained via IV. Online Appendix Figure A.III shows that this close correspondence between model and nonparametric $LATE_h$ holds even across covariate groups with very different average treatment effects. The remaining rows of Table VIII report estimates of average effects for compliers relative to specific care alternatives

TABLE VIII
TREATMENT EFFECTS FOR SUBPOPULATIONS

		Control function		
Parameter	(1) IV	(2) Covariates	(3) Sites	(4) Full model
$LATE_h$	0.247	0.261	0.190	0.214
	(0.031)	(0.032)	(0.076)	(0.042)
$LATE_{nh}$		0.386	0.341	0.370
		(0.143)	(0.219)	(0.088)
$LATE_{ch}$		0.023	−0.122	−0.093
		(0.251)	(0.469)	(0.154)
Lowest predicted quintile:				
$LATE_h$		0.095	0.114	0.027
		(0.061)	(0.112)	(0.067)
$LATE_h$ with fixed S_c		0.125	0.125	0.130
		(0.060)	(0.434)	(0.119)
Highest predicted quintile:				
$LATE_h$		0.402	0.249	0.472
		(0.042)	(0.173)	(0.079)
$LATE_h$ with fixed S_c		0.364	0.289	0.350
		(0.056)	(1.049)	(0.126)

Notes. This table reports estimates of treatment effects for subpopulations. Column (1) reports an IV estimate of the effect of Head Start. Columns (2)–(4) show estimates of treatment effects computed from the control function models displayed in Table VII. The bottom rows show effects in the lowest and highest quintiles of model-predicted LATE. Rows with fixed c-complier shares weight subLATEs using the full-sample estimate of this share (0.34). Standard errors are bootstrapped and clustered at the center level.

(i.e., subLATEs).[18] Estimates of the subLATE for n-compliers, $LATE_{nh}$, are stable across specifications and indicate that the impact of moving from home care to Head Start is large—on the order of 0.37 standard deviations. By contrast, estimates of $LATE_{ch}$, though more variable across specifications, never differ significantly from zero.

Our estimates of $LATE_{nh}$ are somewhat smaller than the average treatment effects of Head Start relative to home care displayed in Table VII. This is a consequence of the reverse Roy pattern captured by the control function coefficients: families willing to switch from home care to Head Start in response to an offer have stronger than average tastes for Head Start, implying smaller than average gains. We can reject that predicted

18. We compute the subLATEs by integrating over the relevant regions of X_i, v_{ih}, and v_{ic} as described in Online Appendix F.

effects of moving from home care to Head Start are equal for n-compliers and n-never takers, implying that this pattern is statistically significant ($p = .038$). Likewise, $LATE_{hc}$ is slightly negative, whereas the average treatment effect of Head Start relative to other preschools is positive ($0.47 - 0.11$). In other words, switching from c to h reduces test scores for c-compliers but would improve the score of an average student. This reflects a combination of above average tastes for competing preschools among c-compliers and positive selection on gains into other preschools. Note that the control function coefficients in Table VII capture selection conditional on covariates and sites, while the treatment effects in Table VIII average over the distribution of observables for each subgroup. The subLATE estimates show that the selection patterns discussed here still hold when variation in effects across covariate and site groups is taken into account.

Another interesting point of comparison is to the 2SLS estimates of Table V. The 2SLS approach found a somewhat smaller $LATE_{nh}$ than our two-step estimator. It also found that Head Start preschools were slightly more effective at raising test scores than were competing programs ($LATE_{ch} > 0$), whereas our full control function estimates suggest the opposite. Importantly, the control function estimates corroborate the failed overidentification tests of Table V by detecting substantial heterogeneity in the underlying subLATEs. This can be seen in the last four rows of Table VIII, which report estimates for the top and bottom quintiles of the model-predicted distribution of $LATE_h$ (see Online Appendix F for details). Fixing each group's S_c at the population average brings estimates for the top and bottom quintiles closer together, but a large gap remains due to subLATE heterogeneity.

Finally, it is worth comparing our findings with those of Feller et al. (forthcoming), who use the principal stratification framework of Frangakis and Rubin (2002) to estimate effects on n- and c-compliers in the HSIS. They also find large effects for compliers drawn from home and negligible effects for compliers drawn from other preschools, though their point estimate of $LATE_{nh}$ is somewhat smaller than ours (0.21 versus 0.37). This difference reflects a combination of different test score outcomes (Feller et al. look only at PPVT scores) and different modeling assumptions. Since neither estimation approach nests the other, it is reassuring that we find qualitatively similar results.

TABLE IX

BENEFITS AND COSTS OF HEAD START WHEN COMPETING PRESCHOOLS ARE RATIONED

(1) Parameter	(2) Description	(3) Value	(4) Source
$LATE_h$	Head Start local average treatment effect	0.247	HSIS
$LATE_{nc}$	Effect of other centers for marginal children	0	Naive assumption: no effect of competing preschools
		0.370	Homogeneity assumption: $LATE_{nc}$ equals $LATE_{nh}$
		0.294	Model-based prediction
NMB	Marginal benefit to Head Start population net of taxes	\$5,513	$(1 - \tau)p(LATE_h + LATE_{nc} \cdot S_c)$, naive assumption
		\$8,321	Homogeneity assumption
		\$7,744	Model-based prediction
MFC	Marginal fiscal cost of Head Start enrollment	\$5,031	$\varphi_h - \tau p(LATE_h + LATE_{nc} \cdot S_c)$, naive assumption
		\$3,519	Homogeneity assumption
		\$3,830	Model-based prediction
MVPF	Marginal value of public funds	1.10	Naive assumption
		2.36	Homogeneity assumption
		2.02	Model-based prediction

Notes. This table reports results of a rate of return calculation for Head Start, assuming that competing preschools are rationed and that marginal students offered seats in these programs as a result of Head Start expansion would otherwise receive home care. Parameter values are obtained from the sources listed in column (4).

IX. Policy Counterfactuals

We now use our model estimates to consider policy counterfactuals that are not nonparametrically identified by the HSIS experiment.

IX.A. Rationed Substitutes

In the cost-benefit analysis of Section VI, we assumed that seats at competing preschools were not rationed. Although this assumption is reasonable in states with universal preschool mandates, other areas may have preschool programs that face relatively fixed budgets and offer any vacated seats to new children. In this case, increases in Head Start enrollment will create opportunities for new children to attend substitute preschools rather than generating cost savings in these programs. Our model-based estimates allow us to assess the sensitivity of our cost-benefit results to the possibility of rationing in competing programs.

From equation (9), the marginal value of public funds under rationing depends on $LATE_{nc}$—the average treatment effect of competing preschools on "*n*- to *c*-compliers" who would move from home care to a competing preschool program in response to an offered seat. We compute the $MVPF_{\delta,rat}$ under three alternative assumptions regarding this parameter. First, we consider the case where $LATE_{nc} = 0$. Next, we consider the case where the average test score effect of competing preschools for marginal students equals the corresponding effect for Head Start compliers drawn from home care (i.e., $LATE_{nc} = LATE_{nh}$). Finally, we use our model to construct an estimate for $LATE_{nc}$. Specifically, we compute average treatment effects of competing preschools relative to home care for students who would be induced to move along this margin by an increase in $U_i(c)$ equal to the utility value of the Head Start offer coefficient.[19] This calculation assumes that the utility value households place on an offered seat

19. Ideally we would compute $LATE_{nc}$ for students who do not receive offers to competing programs but would attend these programs if offered. Because we do not observe offers to substitute preschools, it is not possible to distinguish between nonoffered children and children who decline offers. Our estimate of $LATE_{nc}$ therefore captures a mix of effects for compliers who would respond to offers and children who currently decline offers but would be induced to attend competing programs if these programs became more attractive.

at a competing program is comparable to the value of a Head Start offer.

Table IX shows the results of this analysis. Setting $LATE_{nc}$ = 0 yields an $MVPF_{\delta,rat}$ of 1.10. This replicates the naive analysis with $\varphi_c = 0$ in the nonrationed analysis. Both of these cases ignore costs and benefits due to substitution from competing programs. Assuming that $LATE_{nc} = LATE_{nh}$ produces a benefit-cost ratio of 2.36. Finally, our preferred model estimates from Section VIII predict that $LATE_{nc}$ = 0.294, which produces a ratio of 2.02. These results suggest that under plausible assumptions about the effects of competing programs relative to home care, accounting for the benefits generated by vacated seats in these programs yields estimated social returns larger than those displayed in Table IV, Panel B.

IX.B. Structural Reforms

We next predict the social benefits of a reform that expands Head Start by making it more attractive rather than by extending offers to additional households. This reform is modeled as an improvement in the structural program feature f, as described in Section V. Examples of such reforms might include increases in transportation services, outreach efforts, or spending on other services that make Head Start attractive to parents. Increases in f are assumed to draw additional households into Head Start but to have no effect on potential outcomes, which rules out peer effects generated by changes in student composition. We use the estimates from our preferred model to compute marginal treatment effects and marginal values of public funds for such reforms, treating changes in f as shifts in the mean Head Start utility $\psi_h(X,Z)$.

Figure I, Panel A displays predicted effects of structural reforms on test scores. Since the program feature has no intrinsic scale, the horizontal axis is scaled in terms of the Head Start attendance rate, with a vertical line indicating the current rate (f = 0). The right axis measures $\vec{S}_c$—the share of marginal students drawn from other preschools. The left axis measures test score effects. The figure plots average treatment effects for subgroups of marginal students drawn from home care and other preschools, along with MTE_h, a weighted average of alternative-specific effects.

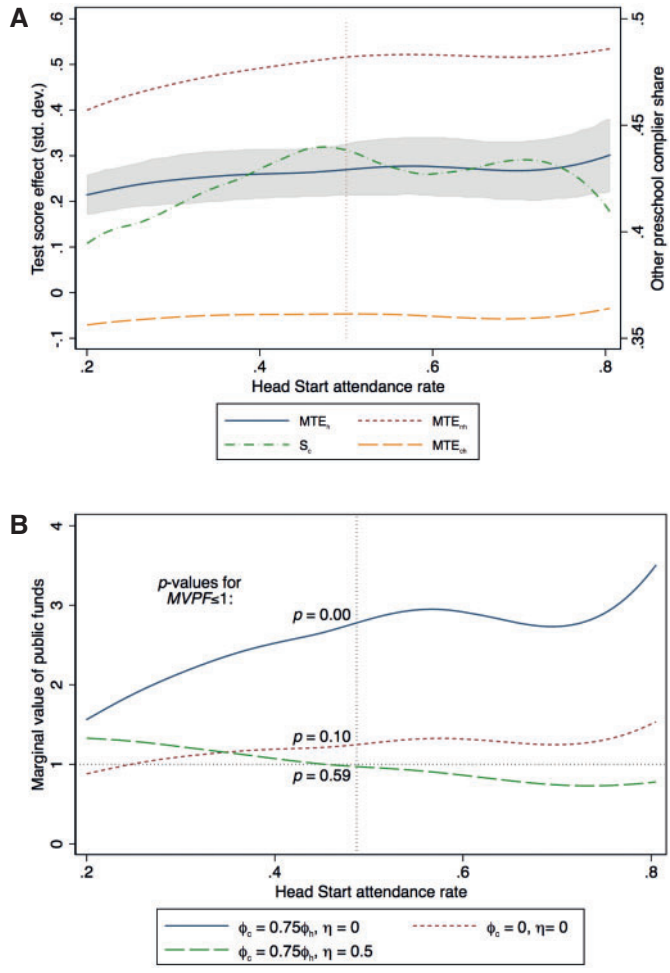

FIGURE I

Effects of Structural Reforms

This figure plots predicted test score effects and marginal values of public funds for various values of the program feature f, which shifts the utility of Head Start attendance. Horizontal axes shows the Head Start attendance rate at each f, and a vertical line indicates the HSIS attendance rate ($f = 0$). Panel A shows marginal treatment effects and competing preschool compliance shares. The left axis measures test score effects. MTE_h is the average effect for marginal students, and MTE_{nh} and MTE_{ch} are effects for subgroups of marginal students drawn from home care and other preschools. The right axis measures the share of marginal students drawn from other preschools. The shaded region shows a 90% symmetric bootstrap confidence interval for MTE_h. Panel B shows predicted marginal values of public funds for structural reforms, using the same parameter calibrations as Table IV. P-values come from bootstrap tests of the hypothesis that the marginal value of public funds is less than or equal to 1 at $f = 0$.

Figure I shows that Head Start's effects on marginal home compliers increase modestly with enrollment and then level out in the neighborhood of the current program scale ($f = 0$). This pattern is driven by reverse Roy selection for children drawn from home care: increases in f attract children with weaker tastes for Head Start, leading to increases in effects for compliers who would otherwise stay home. Predicted effects for children drawn from other preschools are slightly negative for all values of f. At the current program scale, the model predicts that the share of marginal students drawn from other preschools is larger for structural reforms than for an increase in the offer rate (0.44 versus 0.35). This implies that marginal compliers are more likely to be drawn from other preschools than are inframarginal compliers. As a result, the value of MTE_h is comparable to the experimental $LATE_h$, despite very large effects on marginal children drawn from home care (roughly 0.5 standard deviations).

To investigate the consequences of this pattern for the social return to Head Start, Panel B plots $MVPF_f$, the marginal value of public funds for structural reforms. This figure relies on the same parameter calibrations as Table IV. Calculations of $MVPF_f$ must account for the fact that changes in structural program features may increase the direct costs of the program. This effect is captured in equation (12) by the term η, which gives the elasticity of the per child cost of Head Start with respect to the scale of the program. Without specifying the program feature being manipulated, there is no natural value for η. We start with the extreme case where $\eta = 0$, which allows us to characterize costs and benefits associated with reforms that draw in children on the margin without changing the per capita cost of the program. We then consider how the cost-benefit calculus changes when $\eta > 0$.

As in our basic cost-benefit analysis, the results in Figure I, Panel B show that accounting for the public savings associated with program substitution has an important effect on the marginal value of public funds. The short-dashed curve plots $MVPF_f$ setting $\varphi_c = 0$. This calibration suggests a marginal value of public funds slightly above 1 at the current program scale, similar to the naive calibration in Table IV. The solid curve accounts for public savings by setting φ_c equal to our preferred value of $0.75\varphi_h$. This generates an upward shift and steepens the $MVPF_f$ schedule, indicating that both marginal and average social returns increase with program scale. The implied marginal value of public funds at the current program scale ($f = 0$) is above 2. This is larger

than the $MVPF_\delta$ of 1.84 reported in Table IV, which indicates the social returns to marginal expansions that shift the composition of compliers are greater than those for expansions that simply raise the offer rate.

The final scenario in Panel B shows $MVPF_f$ when $\varphi_c = 0.75\varphi_h$ and $\eta = 0.5$.[20] This scenario implies sharply rising marginal costs of Head Start provision: an increase in f that doubles enrollment raises per capita costs by 50%. In this simulation the marginal value of public funds is roughly equal to 1 when $f = 0$, and falls below 1 for higher values. Hence, if η is at least 0.5, a \$1 increase in Head Start spending generated by structural reform will result in less than \$1 transferred to Head Start applicants. This exercise illustrates the quantitative importance of determining provision costs when evaluating specific policy changes such as improvements to transportation services or marketing.

Our analysis of structural reforms suggests increasing returns to the expansion of Head Start in the neighborhood of the current program scale—expansions will draw in households with weaker tastes for preschool with above average potential gains. These findings imply that structural reforms targeting children who are currently unlikely to attend Head Start and children that are likely to be drawn from nonparticipation will generate larger effects than reforms that simply create more seats. Our results also echo other recent studies finding increasing returns to early childhood investments, though the mechanism generating increasing returns in these studies is typically dynamic complementarity in human capital investments rather than selection and effect heterogeneity (see, e.g., Cunha, Heckman, and Schennach 2010).

X. Conclusion

Our analysis suggests that Head Start, in its current incarnation, passes a strict cost-benefit test predicated only on projected effects on adult earnings. It is reasonable to expect that this conclusion would be strengthened by incorporating the value of any impacts on crime (as in Lochner and Moretti 2004 and

20. For this case, marginal costs are obtained by solving the differential equation $\varphi_h'(f) = \eta\varphi(f)\left(\frac{\partial \ln P(D_i = h)}{\partial f}\right)$ with the initial condition $\varphi_h(0) = 8{,}000$. This yields the solution $\varphi_h(f) = 8{,}000\exp\left(\eta(ln\, P(D_i = h) - ln\, P_0)\right)$ where P_0 is the initial Head Start attendance rate.

Heckman et al. 2010) or other externalities such as civic engagement (Milligan, Moretti, and Oreopoulos 2004) or by incorporating the value to parents of subsidized care (as in Aaberge et al. 2010). We find evidence that Head Start generates especially large benefits for children who would not otherwise attend preschool and for children with weak unobserved tastes for the program. This suggests that the program's rate of return can be boosted by reforms that target new populations, though this necessitates the existence of a cost-effective technology for attracting these children.

The finding that returns are on average greater for nonparticipants is potentially informative for the debate over calls for universal preschool, which might reach high-return households. However, it is important to note that if competing state-level preschool programs become ubiquitous, the rationale for expansions to federal preschool programs could be undermined. To see this, consider how the marginal value of expanding Head Start changes as the compliance share S_c approaches 1, so that nearly all denied Head Start applicants would otherwise enroll in competing programs. If Head Start and competing programs have equivalent effects on test scores, then equation (8) indicates that we should decide between federal and state-level provision based entirely on cost criteria. Since state programs are often cheaper (CEA 2014) and are expanding rapidly, the case for federal preschool may actually be weaker now than at the time of the HSIS.

It is important to note some other limitations to our analysis. First, our cost-benefit calculations rely on literature estimates of the link between test score effects and earnings gains. These calculations are necessarily speculative, as the only way to be sure of Head Start's long-run effects is to directly measure long-run outcomes for HSIS participants. Second, we have ignored the possibility that substantial changes to program features or scale could, in equilibrium, change the education production technology. For example, implementing recent proposals for universal preschool could generate a shortage of qualified teachers (Rothstein 2015). Finally, we have ignored the possibility that administrative program costs might change with program scale, choosing instead to equate average with marginal provision costs.

Despite these caveats, our analysis has shown that accounting for program substitution in the HSIS experiment is crucial for an assessment of the Head Start program's costs and benefits.

Similar issues arise in the evaluation of job training programs (Heckman et al. 2000), health insurance (Finkelstein et al. 2012), and housing subsidies (Kling, Liebman, and Katz 2007; Jacob and Ludwig 2012). The tools developed here are potentially applicable to a wide variety of evaluation settings where data on enrollment in competing programs are available.

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at QJE online (qje.oxfordjournals.org).

UNIVERSITY OF CALIFORNIA, BERKELEY, AND NATIONAL BUREAU OF ECONOMIC RESEARCH
UNIVERSITY OF CALIFORNIA, BERKELEY, AND NATIONAL BUREAU OF ECONOMIC RESEARCH

REFERENCES

Aaberge, Rolf, Manudeep Bhuller, Audun Langørgen, and Magne Mogstad, "The Distributional Impact of Public Services When Needs Differ," *Journal of Public Economics*, 94 (2010), 549–562.
Abdulkadiroğlu, Atila, Joshua Angrist, and Parag Pathak, "The Elite Illusion**:** Achievement Effects at Boston and New York Exam Schools," *Econometrica*, 82 (2014), 137–196.
Angrist, Joshua, Guido Imbens, and Donald Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91 (1996), 444–455.
Angrist, Joshua, and Jörn-Steffen Pischke, *Mostly Harmless Econometrics* (Princeton, NJ: Princeton University Press, 2009).
Barnett, W. Steven, "Effectiveness of Early Educational Intervention," *Science*, 333 (2011), 975–978.
Bitler, Marianne, Thurston Domina, and Hilary Hoynes, "Experimental Evidence on Distributional Effects of Head Start," NBER Working Paper 20434, 2014.
Bloom, Howard, and Christina Weiland, "Quantifying Variation in Head Start Effects on Young Children's Cognitive and Socio-Emotional Skills Using Data from the National Head Start Impact Study," MDRC Report, 2015.
Bonhomme, Stéphane, and Elena Manresa, "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83 (2015), 1147–1184.
Bound, John, David Jaeger, and Regina Baker, "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90 (1995), 443–450.
Brinch, Christian, Magne Mogstad, and Matthew Wiswall, "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, forthcoming.
Carneiro, Pedro, and Rita Ginja, "Long-Term Impacts of Compensatory Preschool on Health and Behavior: Evidence from Head Start," *American Economic Journal: Economic Policy*, 6 (2014), 135–173.
Cascio, Elizabeth, and Diane Whitmore Schanzenbach, "The Impacts of Expanding Access to High-Quality Preschool Education," *Brookings Papers on Economic Activity*, Fall (2013), 127–192.

Chetty, Raj, "Sufficient Statistics for Welfare Analysis: A Bridge between Structural and Reduced-Form Methods," *Annual Review of Economics*, 1 (2009), 451–488.
Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics*, 126 (2011), 1593–1660.
Chetty, Raj, John Friedman, and Jonah Rockoff, "Measuring the Impacts of Teachers I: Measuring Bias in Teacher Value-added Estimates," *American Economic Review*, 104 (2014a), 2593–2632.
———, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104 (2014b), 2633–2679.
Congressional Budget Office (CBO), "Effective Marginal Tax Rates for Low- and Moderate-Income Workers," 2012, available at https://www.cbo.gov/sites/default/files/11-15-2012-MarginalTaxRates.pdf.
Council of Economic Advisers (CEA), "The Economics of Early Childhood Investments," 2014, available at https://www.whitehouse.gov/sites/default/files/docs/early_childhood_report1.pdf.
Cox, Nicholas, "Speaking Stata: Correlation with Confidence, or Fisher's Z Revisited," *Stata Journal*, 8 (2008), 413–439.
Cunha, Flavio, James Heckman, and Susanne Schennach, "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78 (2010), 883–931.
Currie, Janet, "Early Childhood Education Programs," *Journal of Economic Perspectives*, 15 (2001), 213–238.
Currie, Janet, and Duncan Thomas, "Does Head Start Make a Difference?," *American Economic Review*, 85 (1995), 341–364.
Deming, David, "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start," *American Economic Journal: Applied Economics*, 1 (2009), 111–134.
Dubin, Jeffrey, and Daniel McFadden, "An Econometric Analysis of Residential Electric Appliance Holdings," *Econometrica*, 52 (1984), 345–362.
Engberg, John, Dennis Epple, Jason Imbrogno, Holger Sieg, and Ron Zimmer, "Evaluating Education Programs That Have Lotteried Admission and Selective Attrition," *Journal of Labor Economics*, 32 (2014), 27–63.
Federal Register, "Executive Order 13330 of February 24, 2004," *Federal Register*, 69 (38) (2004), 9185–9187.
Feller, Avi, Todd Grindal, Luke Miratrix, and Lindsay Page, "Compared to What? Variation in the Impact of Early Childhood Education by Alternative Care-Type Settings," *Annals of Applied Statistics*, forthcoming.
Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group, "The Oregon Health Insurance Experiment: Evidence from the First Year," *Quarterly Journal of Economics*, 127 (2012), 1057–1106.
Frangakis, Constantine, and Donald Rubin, "Principal Stratification in Causal Inference," *Biometrics*, 58 (2002), 21–29.
Garces, Eliana, Duncan Thomas, and Janet Currie, "Longer-Term Effects of Head Start," *American Economic Review*, 92 (2002), 999–1012.
Gelber, Alexander, and Adam Isen, "Children's Schooling and Parents' Investment in Children: Evidence from the Head Start Impact Study," *Journal of Public Economics*, 101 (2013), 25–38.
Geweke, John, "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57 (1989), 1317–1339.
Gibbs, Chloe, Jens Ludwig, and Douglas Miller, "Does Head Start Do Any Lasting Good?," NBER Working Paper 17452, 2011.
Hajivassiliou, Vassilis, and Daniel McFadden, "The Method of Simulated Scores for the Estimation of LDV Models," *Econometrica*, 66 (1998), 863–898.
Hall, Peter, *The Bootstrap and Edgeworth Expansion* (New York: Springer, 1992).
Heckman, James, "Sample Selection Bias as a Specification Error," *Econometrica*, 47 (1979), 153–161.

Heckman, James, Neil Hohmann, Jeffrey Smith, and Michael Khoo, "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment," *Quarterly Journal of Economics*, 115 (2000), 651–694.
Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz, "The Rate of Return to the High/Scope Perry Preschool Program," *Journal of Public Economics*, 94 (2010), 114–128.
Heckman, James, Rodrigo Pinto, and Peter Savelyev, "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, 103 (2013), 2052–2086.
Heckman, James, Sergio Urzua, and Edward Vytlacil, "Instrumental Variables in Models with Multiple Outcomes: The General Unordered Case," *Annales d'Economie et de Statistique*, 91/92 (2008), 151–174.
Heckman, James, and Edward Vytlacil, "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96 (1999), 4730–4734.
Hendren, Nathaniel, "The Policy Elasticity," *Tax Policy and the Economy*, 30 (2016).
Hull, Peter. "IsoLATEing: Identifying Counterfactual-Specific Treatment Effects by Stratified Comparison," working paper, 2015.
Imbens, Guido, and Joshua Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62 (1994), 467–475.
Jacob, Brian, and Jens Ludwig, "The Effects of Housing Assistance on Labor Supply: Evidence from a Voucher Lottery," *American Economic Review*, 102 (2012), 272–304.
Kane, Thomas, Jonah Rockoff, and Douglas Staiger, "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City," *Economics of Education Review*, 27 (2008), 615–631.
Keane, Michael, "A Computationally Practical Simulation Estimator for Panel Data," *Econometrica*, 62 (1994), 95–116.
Kirkeboen, Lars, Edwin Leuven, and Magne Mogstad, "Field of Study, Earnings, and Self-Selection," *Quarterly Journal of Economics*, 131 (2016), doi:10.1093/qje/qjw019 (this issue).
Klein, Joe, "Time to Ax Public Programs that Don't Yield Results," *Time*, July 7, 2011.
Kling, Jeffrey, Jeffrey Liebman, and Lawrence Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 75 (2007), 83–119.
Lafontaine, Francine, and Kenneth White, "Obtaining Any Wald Statistic You Want," *Economics Letters*, 21 (1986), 35–40.
Lee, Chul-In, and Gary Solon, "Trends in Intergenerational Income Mobility," *Review of Economics and Statistics*, 91 (2009), 766–772.
Lochner, Lance, and Enrico Moretti, "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports," *American Economic Review*, 94 (2004), 155–189.
Long, Cuiping, "Experimental Evidence of the Effect of Head Start on Maternal Human Capital Investment," working paper, 2015.
Ludwig, Jens, and Douglas Miller, "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics*, 122 (2007), 159–208.
Ludwig, Jens, and Deborah Phillips, "The Benefits and Costs of Head Start," NBER Working Paper 12973, 2007.
Mayshar, Joram, "On Measures of Excess Burden and Their Application," *Journal of Public Economics*, 43 (1990), 263–289.
Milligan, Kevin, Enrico Moretti, and Philip Oreopoulos, "Does Education Improve Citizenship? Evidence from the United States and the United Kingdom," *Journal of Public Economics*, 88 (2004), 1667–1695.
Puma, Michael, Stephen Bell, Ronna Cook, and Camilla Heid, *Head Start Impact Study: Final Report* (Washington, DC: U.S. Department of Health and Services, Administration for Children and Families, 2010).
Puma, Michael, Stephen Bell, and Camilla Heid, *Third Grade Follow-Up to the Head Start Impact Study* (Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, 2012).

Rothstein, Jesse, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics*, 125 (2010), 175–214.
———, "Teacher Quality Policy When Supply Matters," *American Economic Review*, 105 (2015), 100–130.
Roy, A. D., "Some Thoughts on the Distribution of Earnings," *Oxford Economics Papers*, 3 (1951), 135–146.
Saggio, Raffaele, "Discrete Unobserved Heterogeneity in Discrete Choice Panel Data Models," master's thesis, Center for Monetary and Financial Studies, 2012.
Schumacher, Rachel, Mark Greenberg, and Janellen Duffy, *The Impact of TANF Funding on State Child Care Subsidy Programs* (Washington, DC: Center for Law and Social Policy, 2001).
Stossel, John, "Head Start Has Little Effect by Grade School?" *Fox Business Television*, March 7, 2014.
U.S. Department of Health and Human Services (DHHS), Administration for Children and Families, "Child Care and Development Fund Fact Sheet," 2012, available at http://www.acf.hhs.gov/sites/default/files/occ/ccdf_fact-sheet.pdf.
———, "Head Start Program Facts, Fiscal Year 2013," 2013, available at http://eclkc.ohs.acf.hhs.gov/hslc/mr/factsheets/docs/hs-program-fact-sheet-2011-final.pdf.
———, "Head Start Services," 2014, available at http://www.acf.hhs.gov/programs/ohs/about/head-start.
Walters, Christopher, "The Demand for Effective Charter Schools," NBER Working Paper 20640, 2014.
———, "Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start," *American Economic Journal: Applied Economics*, 7 (2015), 76–102.