NBER WORKING PAPER SERIES

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

Patrick Kline
Christopher Walters

Working Paper 21658
http://www.nber.org/papers/w21658

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2015

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, and Vanderbilt University 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 views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Evaluating Public Programs with Close Substitutes: The Case of Head Start
Patrick Kline and Christopher Walters
NBER Working Paper No. 21658
October 2015, Revised March 2016
JEL No. H43

ABSTRACT

This paper empirically evaluates the cost-effectiveness of Head Start, the largest early-childhood education program in the United States. Using data from the Head Start Impact Study (HSIS), we show that Head Start draws roughly a third of its participants from competing preschool programs, many of which receive public funds. Accounting for the public savings associated with reduced enrollment in other subsidized preschools substantially increases estimates of the program's rate of return. To parse Head Start'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 offer status and household characteristics. We find that Head Start's effects are greater for children who would not otherwise attend preschool and for children that are less likely to participate in the program.

Patrick Kline
Department of Economics
University of California at Berkeley
530 Evans Hall #3880
Berkeley, CA 94720
and NBER
pkline@econ.berkeley.edu

Christopher Walters
Department of Economics
University of California at Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
crwalters@econ.berkeley.edu

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.

This paper 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 3- and 4-year-old children at a cost of $7.6 billion (US 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 et al., 2002; Ludwig and Miller, 2007; Deming, 2009; Carneiro and Ginja, forthcoming). 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, 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 such conclusions are premature. First, research on early childhood interventions finds long run gains in adult outcomes despite short run fadeout of test score impacts (Heckman et al., 2010, 2013; Chetty et al., 2011, 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 paper 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 fadeout is less clear than naive intent-to-treat estimates suggest. Turning to substitution patterns, we find that roughly one third of Head Start compliers 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" (Angrist, Imbens, and Rubin, 1996). 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. While such effects are typically ignored in cost-benefit analyses of Head Start and other similar programs (e.g., Council of Economic Advisers, 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 to 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. (2014), 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 unlikely to undermine the favorable estimated rates of return found in our baseline analysis.

The rest of the paper 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 percent 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 percent 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 (US DHHS, 2014). It is also oversubscribed: in 2002, 85 percent 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 childcare vouchers that can be used for center-based preschool (US DHHS, 2012). Most states also use TANF funds to finance additional childcare subsidies (Schumacher et al., 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 non-experimental 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 non-participant siblings (Currie and Thomas, 1995; Garces et al., 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, forthcoming). These studies show longer run improvements in health outcomes and criminal activity.

In contrast to these non-experimental 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 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 the Fall of 2002. The experimental results showed that a Head Start offer increased measures of cognitive achievement by roughly 0.1 standard deviations during preschool, but that these gains faded out by kindergarten. Moreover, the experiment showed little evidence of effects on non-cognitive or health outcomes (Puma et al., 2010, 2012). These results suggest both smaller short-run effects and faster fadeout than non-experimental estimates for earlier cohorts. Scholars and policymakers 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 non-experimental literature could be due to changes in program effectiveness over time or to selection bias in non-experimental sibling compar-

[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 fadeout. Gelber and Isen (2013) find persistent effects on parental engagement with children. Bitler et al. (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.

isons. Another explanation, however, is that these two research designs identify different parameters. Most non-experimental 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 first 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 non-missing baseline characteristics and Spring 2003 test scores. 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, normed to have mean zero and variance one 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 also available in each year of the experiment, allowing us to produce comparable estimates over time.

Table I provides summary statistics for our analysis sample. The HSIS experiment included two age cohorts: 55 percent of applicants were randomized at age 3 and could attend Head Start for up to two years, while the remaining 45 percent were randomized at age 4 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 percent 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

[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. Because the numbers of treatment and control children at each center were fixed in advance, this is an error-free measure of the probability of an offer (Puma et al., 2010).

[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 Appendix A for details.

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 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 three. Our data includes 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 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, while the intent-to-treat falls roughly in proportion, generating a second-year IV estimate of 0.245 for this cohort. Estimates in years three and four 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

[4]The control vector includes gender, 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. 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.

with a standard error of 0.060. Notably, the 95-percent 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 fadeout 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 now turn 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 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:

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

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 $D_i(0)$:

1. n-compliers: $D_i(1) = h,\ D_i(0) = n$,

2. c-compliers: $D_i(1) = h,\ D_i(0) = c$,

[6] One might also be interested in the effects of Head Start on non-cognitive outcomes, which appear to be important mediators of the effects of early childhood programs in other contexts (Chetty et al., 2011; Heckman et al., 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 non-cognitive outcomes differ significantly from those analyzed in previous studies, however, and it is unclear whether they capture the same skills.

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

3. n-never takers: $D_i(1) = D_i(0) = n$,

4. c-never takers: $D_i(1) = D_i(0) = c$,

5. 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 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 non-offered households. In the first year of the experiment 8.3 percent 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 percent of households are n-never takers. As can be seen in column (4), 13.6 percent 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 percent to 8.3 percent, and reduces the share of children in home care from 55 percent to 9.5 percent. This implies that 23.2 percent of households are c-compliers, and 45.5 percent 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 shall assume that Head Start offers affect test scores only through program participation choices. Under assumption (1), IV identifies a variant of the Local Average Treatment Effect (LATE) of Imbens and Angrist (1994), giving the

average effect of Head Start participation for compliers relative to a mix of program alternatives. Specifically, under (1) and excludability of Head Start offers:

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

The left-hand side of (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:

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

where $LATE_{dh} \equiv E\left[Y_i(h)-Y_i(d)|D_i(1)=h, D_i(0)=d\right]$ 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 Appendix B). In the first year of the HSIS experiment, 34 percent 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 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, while others receive funding from a mix of public sources (see 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). We next 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, forthcoming), 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 Head Start Impact Study find no short term impacts on parents' work decisions (Puma et al., 2010, 2012).[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 upon 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 Local Average Treatment Effect 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 Marginal Treatment Effect 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, a competing preschool program (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 below.

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, while 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). Since offers are assigned at random, market shares for the three care environments can be written $P(D_i = d) =$

[8] We replicate this analysis for our sample in 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- vs. part-time. Recent work by Long (2015) suggests that Head Start may have small effects on full- vs. part-time work for mothers of three-year-olds.

$\delta P(D_i(1) = d) + (1 - \delta) P(D_i(0) = d)$.

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, 2013; Chetty et al., 2011, 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 mean test scores neglects distributional concerns which may lead us to undervalue Head Start's test score impacts (see Bitler et al., 2014).

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

$$C = C_0 + \phi_h P(D_i = h) + \phi_c P(D_i = c) - \tau pE[Y_i], \tag{5}$$

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 $\tau pE[Y_i]$ 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 will adjust for in our empirical work via discounting.

V.C. Changing Offer Probabilities

Consider now 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 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

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

HSIS experiment. Hence, (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. Below we explore how this expression changes when the composition of compliers responds to a policy change.

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

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

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

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

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 $\phi_h - \phi_c S_c$ can be viewed as a local average treatment effect 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, 2014), which we can write:

$$MVPF_\delta \equiv \frac{\partial B/\partial \delta}{\partial C/\partial \delta} = \frac{(1-\tau)\, pLATE_h}{\phi_h - \phi_c S_c - \tau pLATE_h}. \tag{8}$$

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 $\phi_c S_c$) and additional tax revenue generated by higher earnings (captured by $\tau pLATE_h$). As emphasized by Hendren (2014), 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 $MVPF_\delta > 1$ a dollar of government spending can raise the after-tax incomes of children by more than a dollar, which is a robust indicator that program expansions are likely to be welfare improving.

An important lesson of the above 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 above analysis presumes that Head Start expansions yield reductions in the enrollment of competing preschools. However, if competing programs are also over-subscribed, 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 Appendix C that under these assumptions the marginal impact of expanding Head Start becomes:

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

where $LATE_{nc} \equiv E\left[Y_i\left(c\right) - Y_i\left(n\right) | D_i\left(Z_{ih}, 1\right) = c, D_i\left(Z_{ih}, 0\right) = n\right]$. 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:

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

Relative to (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:

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

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 Appendix C). Unfortunately, effects for n-to-c compliers are not nonparametrically identified by the HSIS experiment since 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 which has no impact on test scores. For example, Executive Order #13330, issued by President 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 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 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

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

and $\overrightarrow{S}_c$ gives the share of children on the margin of participating in Head Start who prefer the competing program to preschool non-participation. Following the terminology in Heckman et al. (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 (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 $\overrightarrow{S}_c$ 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

[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.

costs. We suppose that changing program features changes the *average* cost $\phi_h(f)$ of Head Start services, so that the net costs to government of financing preschool are now:

$$C(f) = C_0 + \phi_h(f) P(D_i = h) + \phi_c P(D_i = c) - \tau p E[Y_i], \tag{10}$$

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

$$\begin{aligned} \frac{\partial C(f)/\partial f}{\partial P(D_i = h)/\partial f} = & \underbrace{\phi_h}_{\text{Marginal Provision Cost}} + \underbrace{\frac{\partial \phi_h(f)/\partial f}{\partial \ln P(D_i = h)/\partial f}}_{\text{Inframarginal Provision Cost}} \\ & - \underbrace{\phi_c \overrightarrow{S}_c}_{\text{Public Savings}} - \underbrace{\tau p MTE_h}_{\text{Added Revenue}}. \end{aligned} \tag{11}$$

The first term on the right hand side of (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. And the final term gives the additional tax revenue raised by the boost in the marginal enrollee's human capital.

Letting $\eta \equiv \frac{\partial \ln \phi(f)/\partial f}{\partial \ln P(D_i=h)/\partial f}$ 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:

$$MVPF_f \equiv \frac{\partial B/\partial f}{\partial C(f)/\partial f} = \frac{(1-\tau) p MTE_h}{\phi_h(1+\eta) - \phi_c \overrightarrow{S}_c - \tau p MTE_h}. \tag{12}$$

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 next 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

[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.

from the literature. Calibrated parameters are listed in panel A of Table IV. To be conservative, we deliberately bias our calibrations towards understating Head Start's benefits and overstating its costs in order to arrive at a lower bound rate of return. Further details of the calibration exercise are provided in Appendix D.

Panel B of Table IV 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 one.[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 zero, 50%, or 75% of ϕ_h. Our preferred calibration uses $\phi_c = 0.75\phi_h$, reflecting that fact that roughly 75 percent of competing centers are publicly funded (see Appendix D). Setting $\phi_c = 0$ yields a $MVPF_\delta$ of 1.10. Setting ϕ_c equal to $0.5\phi_h$ and $0.75\phi_h$ 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 one when $\phi_c = 0.5\phi_h$ or $0.75\phi_h$. Notably, our preferred estimate of 1.84 is well above the estimated rates of return of comparable expenditure programs summarized in Hendren (2014, Table 1), 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 one for each value of ϕ_c. When $\phi_c = 0$ the breakeven earnings effect is 9 percent per test score standard deviation, only slightly below our calibrated value of 10 percent. 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 one. Increasing ϕ_c to $0.5\phi_h$ or $0.75\phi_h$ reduces the breakeven earnings effect to 8 percent or 7 percent, 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 percent; 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.

[12]This test is computed by a non-parametric 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 over-cover, while 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 upon the standard delta method approximation.

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 impacts the program's rate of return.

VII.A. Instrumental Variables 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; Abdulkadiroglu et al., 2014). Such approaches can secure identification in a constant effects framework but, as we demonstrate in Appendix E, will typically fail to identify interpretable parameters if the subLATEs themselves vary across the interacting groups (see Kirkeboen et al., 2014 and Hull, 2015 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 8 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 instrumental variables 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; Walters, 2015). Using site

[13] Previous analyses of the HSIS have shown important effect heterogeneity with respect to baseline scores and first language (Bitler et al., 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.

interactions as instruments again yields much more independent variation in Head Start enrollment than in competing preschools.[14] However, now 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 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 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 on) 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 parametrizes 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 non-participation to zero, we assume households have utilities over program alternatives given by:

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

[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 Appendix G for details.

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 (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 upon 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, (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 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 (15) across values of the program offer Z_i as follows:

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

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 – i.e. 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. 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 (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 et al. (2012), 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 upon household covariates.

VII.D. Estimation

To make estimation tractable, we approximate $\psi_h(X, Z)$ and $\psi_c(X)$ with flexible linear functions. The non-separability of $\psi_h(X, Z)$ is captured by linear interactions between Z and the 8 covariates used in our earlier 2SLS analysis. We also allow interactions with the 183 experimental site indicators but, to avoid incidental parameters problems, 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 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 $\text{arctanh}\rho(X) = \frac{1}{2}\ln\left(\frac{1+\rho(X)}{1-\rho(X)}\right)$ 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; Hajivassiliou and McFadden, 1998; Keane, 1994). Models including site groups are estimated with an algorithm that alternates between maximizing the likelihood and reassigning groups, described in detail in Appendix G. Second, we use the parameters of the

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

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

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 8 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 Appendix Table A.VI), with group shares that vary between 12% and 21% of the sample. These assignments comprise 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 homogenous. Panel A of Column (2) indicates that the effects of an offer are greater at high-quality centers and lower among non-poor children that would typically be ineligible for Head Start enrollment.[16] Panel B of 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. While there is some evidence of heterogeneity in the correlation based upon 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

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

[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.

in predicted enrollment shares for Head Start and competing preschools. 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 (16), the model is under-identified 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 $\lambda_c(X_i, 1, d) - \lambda_c(X_i, 0, d)$. 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 (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, while the corresponding difference for students in competing preschools is 0.26 standard deviations. Column (2) adds controls for baseline 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 et al., 2008; Rothstein, 2010; Chetty et al., 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.

While 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 positive and in the full model is significant. A possible interpretation of these patterns is that Head Start is viewed by parents as a preschool of last resort, leading to

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

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 Appendix F for some additional goodness of fit tests). While 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. Appendix Figure A.III shows that this close correspondence between model and non-parametric $LATE_h$ holds even across different covariate groups, across which there is enormous heterogeneity. The remaining rows of Table VIII report estimates of average effects for compliers relative to specific care alternatives (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 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 ($p = 0.038$). Likewise, $LATE_{hc}$ is slightly negative, while 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

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

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 above 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 competing programs ($LATE_{ch} > 0$), while 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 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. (2014), 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 vs. 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.

IX. Policy Counterfactuals

We now use our model estimates to consider policy counterfactuals that are not non-parametrically 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. While 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 (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 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 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 $\phi_c = 0$ in the non-rationed 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 panel B of Table IV.

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 $\psi_h(X, Z)$.

Panel A of Figure I 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 $\overrightarrow{S}_c$ – the share of marginal students drawn from other preschools. The left axis measures test

[19] Ideally we would compute $LATE_{nc}$ for students who do not receive offers to competing programs but would attend these programs if offered. Since we do not observe offers to substitute preschools, it is not possible to distinguish between non-offered 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.

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 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 vs. 0.35). This implies that marginal compliers are more likely to be drawn from other preschools than 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 0.5σ).

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 (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 panel B of Figure I show that accounting for the public savings associated with program substitution has an important effect on the marginal value of public funds. The red curve plots $MVPF_f$ setting $\phi_c = 0$. This calibration suggests a marginal value of public funds slightly above one at the current program scale, similar to the naive calibration in Table IV. The blue curve accounts for public savings by setting ϕ_c equal to our preferred value of $0.75\phi_h$. 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 $\phi_c = 0.75\phi_h$ and $\eta = 0.5$.[20] This scenario implies sharply rising marginal costs of Head Start provision: an increase in f that doubles enroll-

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

ment raises per-capita costs by 50 percent. In this simulation the marginal value of public funds is roughly equal to one when $f = 0$, and falls below one for higher values. Hence, if η is at least 0.5, a dollar increase in Head Start spending generated by structural reform will result in less than one dollar 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 non-participation 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 et al., 2010).

X. Conclusion

Our analysis suggests that Head Start, in its current incarnation, passes a strict cost-benefit test predicated only upon 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 (e.g. as in Lochner and Moretti, 2004 and Heckman et al., 2010), or other externalities such as civic engagement (Milligan et al., 2004), or by incorporating the value to parents of subsidized care (e.g., 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 one, so that nearly all denied Head Start applicants would otherwise enroll in competing programs. If Head Start and competing program have equivalent effects on test scores, then (8) indicates that we should decide between federal and state level provision based entirely on cost criteria. Since state programs are often cheaper (Council of Economic Advisers, 2015) and are expanding rapidly, the case for federal preschool may actually be weaker now than at the time of the Head Start Impact Study.

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, forthcoming). 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 et al., 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.

References

1. Aaberge, R., Bhuller, M., Langørgen, A., and Mogstad, M. (2010). "The Distributional Impact of Public Services When Needs Differ." Journal of Public Economics 94(9).

2. Abdulkadiroglu, A., Angrist, J., and Pathak, P. (2014). "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools." Econometrica 82(1).

3. Angrist, J., Imbens, G., and Rubin, D. (1996). "Identification of Causal Effects using Instrumental Variables." Journal of the American Statistical Association 91(434).

4. Angrist, J., and Pischke, S. (2009). Mostly Harmless Econometrics. Princeton, NJ: Princeton University Press.

5. Barnett, W. (2011). "Effectiveness of Early Educational Intervention." Science 333(6045).

6. Bitler, M., Domina, T., and Hoynes, H. (2014). "Experimental Evidence on Distributional Effects of Head Start." NBER Working Paper no. 20434.

7. Bloom, H., and Weiland, C. (2015). "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.

8. Bonhomme, S., and Manresa, E. (2015). "Grouped Patterns of Heterogeneity in Panel Data." Econometrica 83(3).

9. Bound, J., Jaeger, D. A., and Baker, R. M. (1995). "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(430).

10. Brinch, C., Mogstad, M., and Wiswall, M. (2012). "Beyond LATE With a Discrete Instrument: Heterogeneity in the Quantity-Quality Interaction of Children." Working paper.

11. Carneiro, P., and Ginja, R. (forthcoming). "Long-Term Impacts of Compensatory Preschool on Health and Behavior: Evidence from Head Start." American Economic Journal: Economic Policy.

12. Cascio, E., and Schanzenbach, E. (2013). "The Impacts of Expanding Access to High-Quality Preschool Education." Brookings Papers on Economic Activity, Fall 2013.

13. Chetty, R. (2009). "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-form Methods." Annual Review of Economics 1.

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

15. Chetty, R., Friedman, J., and Rockoff, J. (2014a). "Measuring the Impacts of Teachers I: Measuring Bias in Teacher Value-added Estimates." American Economic Review 104(9).

16. Chetty, R., Friedman, J., and Rockoff, J. (2014b). "Measuring the Impacts of Teachers II: Teacher Value-added and Student Outcomes in Adulthood." American Economic Review 104(9).

17. Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014c). "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." Quarterly Journal of Economics 129(4).

18. Congressional Budget Office (2012). "Effective Marginal Tax Rates for Low- and Moderate-Income Workers." https://www.cbo.gov/sites/default/files/11-15-2012-MarginalTaxRates.pdf.

19. Council of Economic Advisers (2015). "The Economics of Early Childhood Investments." Report Prepared by the Executive Office of the President of the United States.

20. Cox, N. (2008). "Speaking Stata: Correlation with Confidence, or Fisher's Z Revisited." The Stata Journal 8(3).

21. Cunha, F., Heckman, J., and Schennach, S. (2010). "Estimating the Technology of Cognitive and Non-cognitive Skill Formation." Econometrica 78(3).

22. Currie, J. (2001). "Early Childhood Education Programs." Journal of Economic Perspectives 15(2).

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

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

25. Dubin, J., and McFadden, D. (1984). "An Econometric Analysis of Residential Electric Appliance Holdings." Econometrica 52 (2).

26. Engberg, J., Epple, D., Imbrogno, J., Sieg, H., and Zimmer, R. (2014). "Evaluating Education Programs That Have Lotteried Admission and Selective Attrition." Journal of Labor Economics 32(1).

27. Federal Register (2004). "Executive Order 13330 of February 24, 2004" 69 (38), 9185-9187.

28. Feller, A., Grindal, T., Miratrix, L., and Page, L. (2014). "Compared to What? Variation in the Impact of Early Childhood Education by Alternative Care-Type Settings." Working paper.

29. Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J., Allen, H., Baicker, K., and the Oregon Health Study Group (2012). "The Oregon Health Insurance Experiment: Evidence from the First Year." Quarterly Journal of Economics 127(3).

30. Frangakis, C., and Rubin, D. (2002). "Principal Stratification in Causal Inference." Biometrics 58(1).

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

32. Gelber, A., and Isen, A. (2013). "Children's Schooling and Parents' Investment in Children: Evidence from the Head Start Impact Study." Journal of Public Economics 101.

33. Geweke, J. (1989). "Bayesian Inference in Econometric Models Using Monte Carlo Integration." Econometrica 57.

34. Gibbs, C., Ludwig, J., and Miller, D. (2011). "Does Head Start Do Any Lasting Good?" NBER Working Paper no. 17452.

35. Hajivassiliou, V., and McFadden, D. (1998). "The Method of Simulated Scores for the Estimation of LDV Models." Econometrica 66.

36. Hall, P. (1992). The Bootstrap and Edgeworth Expansion. Springer: New York, NY.

37. Heckman, J. (1979). "Sample Selection Bias as a Specification Error." Econometrica 47(1).

38. Heckman, J., and Vytlacil, E. (1999). "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." Proceedings of the National Academy of Sciences 96(8).

39. Heckman, J., Hohmann, N., Smith, J., and Khoo, M. (2000). "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment." Quarterly Journal of Economics 115 (2).

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

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

42. Heckman, J., Urzua, S., and Vytlacil, E. (2008). "Instrumental Variables in Models With Multiple Outcomes: The General Unordered Case." Annales d'Economie et de Statistique, 91/92.

43. Hendren, N. (2014). "The Policy Elasticity." Mimeo, Harvard University.

44. Hull, P. (2015). "IsoLATEing: Identifying Counterfactual-Specific Treatment Effects by Stratified Comparisons." Working Paper.

45. Imbens, G., and Angrist, J. (1994). "Identification and Estimation of Local Average Treatment Effects." Econometrica 62.

46. Jacob, B., and Ludwig, J. (2012). "The Effects of Housing Assistance on Labor Supply: Evidence from a Voucher Lottery." American Economic Review 102(1).

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

48. Keane, M. (1994). "A Computationally Practical Simulation Estimator for Panel Data." Econometrica 62.

49. Kirkeboen, L., Leuven, E., and Mogstad, M. (2014). "Field of Study, Earnings, and Self-Selection." Working Paper.

50. Klein, J. (2011). "Time to Ax Public Programs That Don't Yield Results." Time Magazine. `http://content.time.com/time/nation/article/0,8599,2081778,00.html`.

51. Kling, J., Liebman, J., and Katz, L. (2007). "Experimental Analysis of Neighborhood Effects." Econometrica 75.

52. Lafontaine, F., and White, K. (1986). "Obtaining any Wald Statistic you Want." Economics Letters 21(1).

53. Lee, C., and Solon, G. (2009). "Trends in Intergenerational Income Mobility." The Review of Economics and Statistics 91(4).

54. Lochner, L., and Moretti, E. (2004). "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports." American Economic Review 94(1).

55. Long, Cuiping (2015). "Experimental Evidence of the Effect of Head Start on Maternal Human Capital Investment." Working paper.

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

57. Ludwig, J., and Phillips, D. (2007). "The Benefits and Costs of Head Start." NBER Working Paper no. 12973.

58. Mayshar, J. (1990). "On Measures of Excess Burden and Their Application." Journal of Public Economics 43(3).

59. Milligan, K., Moretti, E., and Oreopoulos, P. (2004). "Does Education Improve Citizenship? Evidence from the United States and the United Kingdom." Journal of Public Economics 88(9).

60. Noss, A. (2014). "Household Income: 2013." American Community Survey Briefs.

61. Puma, M., Bell, S., Cook, R., and Heid, C. (2010). "Head Start Impact Study: Final Report." U.S. Department of Health and Services. Administration for Children and Families. Washington, DC.

62. Puma, M., Bell, S., and Heid, C. (2012). "Third Grade Follow-up to the Head Start Impact Study." U.S. Department of Health and Human Services. Washington, DC.

63. Rothstein, J. (2010). "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." Quarterly Journal of Economics 125(1).

64. Rothstein, J. (forthcoming). "Teacher Quality Policy When Supply Matters." American Economic Review.

65. Roy, A. (1951). "Some Thoughts on the Distribution of Earnings." Oxford Economics Papers 3(2).

66. Saggio, R. (2012). "Discrete Unobserved Heterogeneity in Discrete Choice Panel Data Models." Master's Thesis, Center for Monetary and Financial Studies.

67. Schumacher, R., Greenberg, M., and Duffy, J. (2001). "The Impact of TANF Funding on State Child Care Subsidy Programs." Center for Law and Social Policy.

68. Stossel, J. (2014). "Head Start Has Little Effect by Grade School?" Fox Business, March 7th, 2014. Television.

69. US Department of Health and Human Services, Administration for Children and Families (2012). "Child Care and Development Fund Fact Sheet." `http://www.acf.hhs.gov/sites/default/files/occ/ccdf_factsheet.pdf`.

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

71. US Department of Health and Human Services, Administration for Children and Families (2014). "Head Start Services." http://www.acf.hhs.gov/programs/ohs/about/head-start.

72. Walters, C. (2015). "Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start." American Economic Journal: Applied Economics 7(4).

73. Walters, C. (2014). "The Demand for Effective Charter Schools." Working Paper

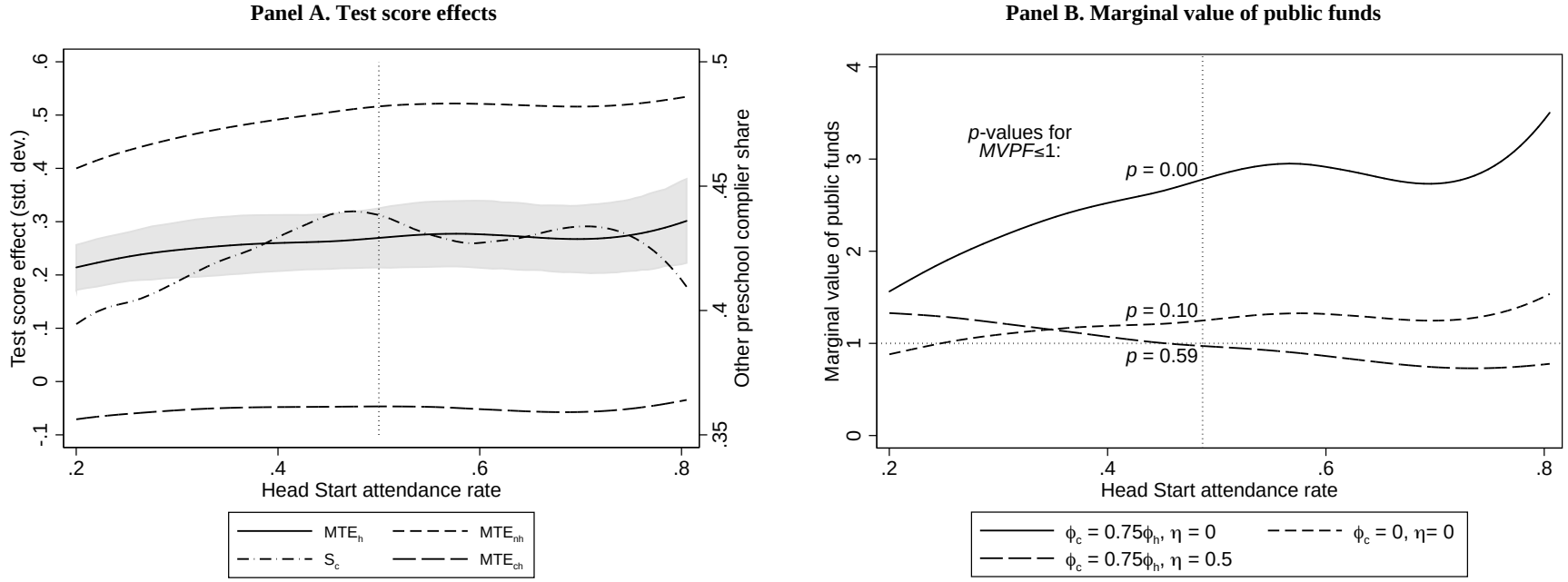

Figure I. Effects of Structural Reforms

Notes: 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, while 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-percent 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 one at $f = 0$.

Table I. Descriptive Statistics

	By offer status			By preschool choice		
Variable	Offered mean (1)	Non-offered mean (2)	Differential (3)	Head Start (4)	Other centers (5)	No preschool (6)
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)			
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	2256	1315	3571	2043	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 percent of observations. Missing values are excluded in statistics for income. The joint *p* -value is from a test of the hypothesis that all coefficients equal zero.

Table II. Experimental Impacts on Test Scores

	Three-year-old cohort			Four-year-old cohort			Cohorts pooled		
	Reduced form	First stage	IV	Reduced form	First stage	IV	Reduced form	First stage	IV
Time period	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
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		1970			1601			3571	
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		1760			1416			3176	
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		1659			1336			2995	
Year 4	0.038	0.344	0.110		-			-	
	(0.034)	(0.029)	(0.098)						
N		1599							

Notes: This table reports experimental estimates of the effects of Head Start on a summary index of test scores. 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 one 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 zero, and dummies for missing are included. Standard errors are clustered by center of random assignment.

Table III. Preschool Choices by Year, Cohort, and Offer Status

		Offered			Not offered			Other center
		Head Start	Other centers	No preschool	Head Start	Other centers	No preschool	complier share
Time period	Cohort	(1)	(2)	(3)	(4)	(5)	(6)	(7)
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 non-offered 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.

Table IV. Benefits and Costs of Head Start

Parameter	Description	Value	Source
(1)	(2)	(3)	(4)
	Panel A. Parameter values		
p	Effect of a 1 SD increase in test scores on earnings	$0.1\bar{e}$	Table A.IV
e_{US}	US 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 US 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 SD 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	Naïve assumption: $\phi_c = 0$
		\$4,000	Pessimistic assumption: $\phi_c = 0.5\phi_h$
		\$6,000	Preferred assumption: $\phi_c = 0.75\phi_h$
	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	$\phi_h - \phi_c S_c - \tau pLATE_h$, naïve assumption
		\$3,671	Pessimistic assumption
		\$2,991	Preferred assumption
$MVPF$	Marginal value of public funds	1.10 (0.22) p-value = 0.1 Breakeven $p/\bar{e} = 0.09$ (0.01)	NMB/MFC (s.e.), naïve assumption
		1.50 (0.34) p-value = 0.00 Breakeven $p/\bar{e} = 0.08$ (0.01)	Pessimistic assumption
		1.84 (0.47) p-value = 0.00 Breakeven $p/\bar{e} = 0.07$ (0.01)	Preferred assumption

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 one. 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 one.

Table V. Two Stage Least Squares Estimates with Interaction Instruments

	One endogenous variable	Two endogenous variables	
	Head Start	Head Start	Other centers
Instruments	(1)	(2)	(3)
Offer	0.247	-	-
(1 instrument)	(0.031)		
Offer x 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	0.007	0.006	
Offer x 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	0.002	0.002	
Offer x 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	0.077	0.050	
Offer x covariates and	0.229	0.302	0.225
offer x site groups (14 instruments)	(0.029)	(0.054)	(0.134)
First-stage *F*	340.2	121.2	13.3
Overid. *p*-value	0.012	0.010	

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 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/Pischke (2009) partial *F*'s. Standard errors are clustered at the center level.

Table VI. Multinomial Probit Estimates

	Head Start utility			
	Main effect	Offer interaction	Other center utility	Arctanh ρ
	(1)	(2)	(3)	(4)
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	0.015	0.000	0.000	0.666
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	0.000	0.000	0.000	0.000

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 Appendix G. The Head Start and other center utilities also include the main effects of gender, 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 zero. Standard errors are clustered at the Head Start center level.

Table VII. Selection-corrected Estimates of Preschool Effects

	Least squares		Control function		
	No controls	Covariates	Covariates	Site groups	Full model
	(1)	(2)	(3)	(4)	(5)
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 x λ_h			-0.167	-0.137	-0.158
			(0.080)	(0.126)	(0.091)
Other preschools x 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 x λ_c			0.224	0.065	0.131
			(0.306)	(0.453)	(0.172)
Other preschools x λ_c			0.488	0.440	0.486
			(0.248)	(0.926)	(0.197)
P-values:					
No selection			0.016	0.510	0.046
No selection on gains			0.133	0.560	0.084
Additive separability			0.261	0.452	0.349

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 gender, 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 de-meaned 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 (3) 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 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 zero in each preschool alternative (see Appendix F for details).

Table VIII. Treatment Effects for Subpopulations

		Control function		
	IV	Covariates	Sites	Full model
Parameter	(1)	(2)	(3)	(4)
$LATE_h$	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 boostrapped and clustered at the center level.

Table IX. Benefits and Costs of Head Start when Competing Preschools are Rationed

Parameter	Description	Value	Source
(1)	(2)	(3)	(4)
$LATE_h$	Head Start Local Average Treatment Effect	0.247	HSIS
$LATE_{nc}$	Effect of other centers for marginal children	0	Naïve assumption: No effect of competing preschools
		0.370	Homogeneity assumption: $n->c$ subLATE equals $n->h$ subLATE
		0.294	Model-based prediction
NMB	Marginal benefit to Head Start population net of taxes	\$5,513	$(1 - \tau)p\,(LATE_h + S_c LATE_{nc})$, naïve assumption
		\$8,321	Homogeneity assumption
		\$7,744	Model-based prediction
MFC	Marginal fiscal cost of Head Start enrollment	\$5,031	$\phi_h - \tau p\,(LATE_h + S_c LATE_{nc})$, naïve assumption
		\$3,519	Homogeneity assumption
		\$3,830	Model-based prediction
MVPF	Marginal value of public funds	1.10	Naïve 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).

Online Appendix

Appendix A: Data

This appendix describes the construction of the sample used in this article. The data come from the Head Start Impact Study (HSIS). This data set includes information on 4,442 children, each applying to Head Start at one of 353 experimental sites in Fall 2002. The raw data used here includes information on test scores, child demographics, preschool attendance, and preschool characteristics. Our core sample includes 3,571 children (80 percent of experimental participants) with non-missing values for key variables. We next describe the procedures used to process the raw data and construct this sample.

Test scores

Outcomes are derived from a series of tests given to students in the Fall of 2002 and each subsequent Spring. The followup window extends through Spring 2006 for the three-year-old applicant cohort and Spring 2005 for the four-year-old cohort.

We use these assessments to construct summary indices of cognitive skills in each period. These summary indices include scores on the Peabody Picture and Vocabulary Test (PPVT) and Woodcock Johnson III Preacademic Skills (WJIII) tests. The WJIII Preacademic Skills score combines performance on several subtests to compute a composite measure of cognitive performance. We use versions of the PPVT and WJIII scores derived from item response theory (IRT), which uses the reliability of individual test items to construct more a more accurate measure of student ability than the simple raw score. The summary index in each period is a simple average of standardized PPVT and WJIII scores, with each score standardized to have mean zero and standard deviation one in the control group, separately by applicant cohort and year. Our core sample excludes applicants without PPVT and WJIII scores in Spring 2003.

The HSIS data includes a number of other test scores in addition to the PPVT and WJIII. Previous analyses of the HSIS data have looked at different combinations of outcomes: Puma et al. (2010) show estimates for each individual test, Walters (2015) uses a summary index that combines all available tests, and Bitler et al. (2014) show separate results for the PPVT and WJIII. We focus on a summary index of the PPVT and WJIII because these tests are among the most reliable in the HSIS data (Puma et al., 2010), are consistently measured in each year (which allows for interpretable intertemporal comparisons), and can be most easily compared to the previous literature (for example, Currie and Thomas, 1995 estimate effects on PPVT scores). Estimates that include additional outcomes in the summary index or restrict attention to individual outcomes produced similar results, though these estimates were typically less precise.

Demographics

Baseline demographics come from a parental survey conducted in Fall 2002. Parents of eighty-one percent of children responded to this survey. We supplement this information with a set of variables in the HSIS "Covariates and Subgroups" data file, which includes additional data collected during experimental recruitment to fill in characteristics for non-respondents. When a characteristic is measured in both files and answers are inconsistent, the "Covariates and Subgroups" value is used. Our core sample excludes applicants with missing values for baseline covariates except income, which is missing more often than other variables. We retain children with missing income and include a missing dummy in all specifications.

Preschool attendance

Preschool attendance is measured from the HSIS "focal arrangement type" variable, which reconciles information from parent interviews and teacher/care provider interviews to construct a summary measure of the childcare setting. This variable includes codes for centers, non-relative's homes, relative's homes, own home (with a relative or non-relative), parent care, and Head Start. Children are coded as attending Head Start if this variable is coded "Head Start;" another preschool center if it is coded "Center;" and no preschool if it takes any other non-missing value. We exclude children with missing focal arrangement types in constructing the core sample.

Preschool characteristics

Our analysis uses experimental site characteristics and characteristics of the preschools children attend (if any), such as whether transportation is provided, funding sources, and an index of quality. This information is derived from interviews with childcare center directors conducted in the Spring of 2003. This information is provided in a student-level file, with the responses of the director of a child's preschool center included as variables. Site characteristics are coded using values of these variables for treatment group children with focal care arrangements coded as "Head Start" at each center of random assignment. In a few cases, these values differed for Head Start attendees at the same site; we used the most frequently-given responses in these cases. An exception is the quality index, which synthesizes information from parent, center director, and teacher surveys. We use the mean value of this index reported by Head Start attendees at each site to construct site-specific measures of quality.

Weights

The probability of assignment to Head Start differed across experimental sites. The HSIS data includes several weight variables designed to account for these differences. These weights also include a factor that adjusts for differences in the probability that Head Start centers themselves were sampled (Puma et al., 2010). This weighting can be used to estimate the average effect of Head Start participation in the US, rather than the average effect in the sample; these parameters

may differ if effects differ across sites in a manner related to sampling probabilities. Probabilities of sampling differed widely across centers, however, leading to very large differences in weights across children and decreasing precision. Instead of using the HSIS weights, we constructed inverse probability weights based on the fraction of applicants at each site offered Head Start. The discussion in Puma et al. (2010) suggests that the numbers of treated and control students at each site were specified in advance, implying that this fraction correctly measures the *ex ante* probability that a child is assigned to the treatment group. Results using other weighting schemes were similar, but less precise.

We also experimented with models including center fixed effects rather than using weights. These models produced similar results, but our multinomial probit model is much more difficult to estimate with fixed effects than with weights. We therefore opted to use weights rather than fixed effects for all estimates reported in the article.

Appendix B: Identification of Complier Characteristics

This appendix extends results from Abadie (2002) to show identification of population shares, characteristics and marginal potential outcome distributions for subpopulations of compliers drawn from other preschools and no preschool. Under the monotonicity restriction (1), we have

$$-\frac{E\left[1\left\{D_i=c\right\}|Z_i=1\right]-E\left[1\left\{D_i=c\right\}|Z_i=0\right]}{E\left[1\left\{D_i=h\right\}|Z_i=1\right]-E\left[1\left\{D_i=h\right\}|Z_i=0\right]}=-\frac{-E\left[1\left\{D_i(0)=c\right\}-1\left\{D_i(1)=c\right\}\right]}{E\left[1\left\{D_i(1)=h\right\}-1\left\{D_i(0)=h\right\}\right]}$$

$$=-\frac{-P\left(D_i(1)=h,D_i(0)=c\right)}{P\left(D_i(1)=h,D_i(0)\neq h\right)}$$

$$=S_c.$$

The share of compliers drawn from competing preschools can therefore be estimated as minus the ratio of the Head Start offer's effect on other preschool attendance to its effet on Head Start attendance.

Observed characteristics and marginal potential outcome distributions for complier subgroups are also identified. Let $g\left(Y_i,X_i\right)$ be any measurable function of outcomes and exogenous covariates. Consider the quantity

$$\kappa_c\equiv\frac{E\left[g(Y_i,X_i)\cdot 1\left\{D_i=c\right\}|Z_i=1\right]-E\left[g\left(Y_i,X_i\right)\right)\cdot 1\left\{D_i=c\right\}|Z_i=0]}{E\left[1\left\{D_i=c\right\}|Z_i=1\right]-E\left[1\left\{D_i=c\right\}|Z_i=0\right]}.$$

The numerator can be written

$$E\left[g\left(Y_i(D_i(1)),X_i\right)\cdot 1\left\{D_i(1)=c\right\}\right]-E\left[g(Y_i(D_i(0)),X_i)\cdot 1\left\{D_i(0)=c\right\}\right],$$

where the conditioning on Z_i has been dropped because offers are independent of potential outcomes and covariates. This simplifies to

$$\kappa_c=E\left[g\left(Y_i(c),X_i\right)|D_i(1)=c\right]P\left(D_i(1)=c\right)-E\left[g\left(Y_i(c),X_i\right)|D_i(0)=c\right]P\left(D_i(0)=c\right)$$

$$=E\left[g\left(Y_i(c),X_i\right)|D_i(1)=c,D_i(0)=c\right]P\left(D_i(1)=c,D_i(0)=c\right)$$

$$-E\left[g\left(Y_i(c),X_i\right)|D_i(1)=c,D_i(0)=c\right]P\left(D_i(1)=c,D_i(0)=c\right)$$

$$-E\left[g\left(Y_i(c),X_i\right)|D_i(1)=h,D_i(0)=c\right]P\left(D_i(1)=h,D_i(0)=c\right)$$

$$=-E\left[g\left(Y_i(c),X_i\right)|D_i(1)=h,D_i(0)=c\right]P\left(D_i(1)=h,D_i(0)=c\right),$$

where the first equality uses the fact that $P\left(D_i(0)=c|D_i(1)=c\right)=1$. The denominator is the effect of the offer on the probability that $D_i=c$, which is minus the share of the population shifted from c to h, $-P\left(D_i(1)=h,D_i(0)=c\right)$. Hence,

$$\kappa_c=\frac{-E\left[g\left(Y_i(c),X_i\right)|D_i(1)=h,D_i(0)=c\right]P\left(D_i(1)=h,D_i(0)=c\right)}{-P\left(D_i(1)=h,D_i(0)=c\right)}$$

$$= E\left[g\left(Y_i(c), X_i\right) | D_i(1) = h, D_i(0) = c\right],$$

which completes the proof.

An analogous argument shows identification of $E\left[g\left(Y_i(n), X_i\right) | D_i(1) = h, D_i(0) = n\right]$ by replacing c with n throughout. Moreover, replacing c with h, the same argument shows identification of $E\left[g\left(Y_i(h), X_i\right) | D_i(1) = h, D_i(0) \neq h\right]$, which can be used to characterize the distribution of $Y_i(h)$ for the full population of compliers.

Note that κ_c is the population coefficient from an instrumental variables regression of $g(Y_i, X_i) \cdot 1\{D_i = c\}$ on $1\{D_i = c\}$, instrumenting with Z_i. The characteristics of the population of compliers shifted from c to h can therefore be estimated using the sample analogue of this regression. In Appendix Table A.II we estimate the characteristics of non-Head Start preschool centers attended by compliers drawn from c by setting $g\left(Y_i, X_i\right)$ equal to a characteristic of the preschool center a child attends (set to zero for children not in preschool). In Appendix Table A.VII we set $g(Y_i, X_i) = Y_i$ to estimate the means of $Y_i(c)$, $Y_i(n)$, and $Y_i(h)$ for compliers.

Appendix C: Derivation of Marginal Value of Public Funds

This appendix derives the expressions for the marginal value of public funds in equations (8), (9) and (12). Section C.4 discusses the use of earnings vs. wage changes to value test score impacts.

C.1 Program Scale

First, consider the case where competing programs are not rationed. From (4), the effect of a change in δ on the average after-tax lifetime income of children is

$$\frac{\partial B}{\partial \delta} = (1-\tau)p\frac{\partial E\left[Y_i\right]}{\partial \delta}.$$

The test score for child i can be written

$$Y_i = Y_i(D_i(1))Z_i + Y_i(D_i(0))(1-Z_i),$$

so

$$E\left[Y_i\right] = E\left[Y_i(D_i(1))|Z_i=1\right]\delta + E\left[Y_i(D_i(0))|Z_i=0\right](1-\delta)$$

$$= E\left[Y_i(D_i(1))\right]\delta + E\left[Y_i(D_i(0))\right](1-\delta),$$

where the second line follows from the assumption that Head Start offers are independent of potential outcomes and potential treatment choices. Then

$$\frac{\partial E\left[Y_i\right]}{\partial \delta} = E\left[Y_i(D_i(1))\right] - E\left[Y_i(D_i(0))\right]$$

$$= E\left[Y_i(D_i(1)) - Y_i(D_i(0))\right]$$

$$= E\left[Y_i(D_i(1)) - Y_i(D_i(0))|D_i(1)\neq D_i(0)\right]P\left(D_i(1)\neq D_i(0)\right).$$

Since $U_i(n)$ and $U_i(c)$ do not depend on Z_i and $U_i(h,1) > U_i(h,0)$, the condition $D_i(1)\neq D_i(0)$ implies that $D_i(1)=h$. We can therefore rewrite the last expression as

$$\frac{\partial E\left[Y_i\right]}{\partial \delta} = E\left[Y_i(h) - Y_i(D_i(0))|D_i(1)=h, D_i(0)\neq h\right]P\left(D_i(1)=h, D_i(0)\neq h\right)$$

$$= LATE_h \cdot P\left(D_i(1)=h, D_i(0)\neq h\right),$$

which is equation (6). It follows that

$$\frac{\partial B}{\partial \delta} = (1-\tau)p\cdot LATE_h \cdot P\left(D_i(1)=h, D_i(0)\neq h\right).$$

From equation (5), the effect of a change in δ on the government budget is

$$\frac{\partial C}{\partial \delta} = \phi_h\frac{\partial P\left(D_i=h\right)}{\partial \delta} + \phi_c\frac{\partial P\left(D_i=c\right)}{\partial \delta} - \tau p\frac{\partial E\left[Y_i\right]}{\partial \delta}.$$

The probability of Head Start participation is

$$P\left(D_i = h\right) = E\left[1\left\{D_i(1) = h\right\}\right]\delta + E\left[1\left\{D_i(0) = h\right\}\right](1-\delta),$$

which implies

$$\begin{aligned}
\frac{\partial P\left(D_i = h\right)}{\partial \delta} &= E\left[1\left\{D_i(1) = h\right\}\right] - E\left[1\left\{D_i(0) = h\right\}\right] \\
&= E\left[1\left\{D_i(1) = h\right\} - 1\left\{D_i(0) = h\right\}\right] \\
&= E\left[1\left\{D_i(1) = h, D_i(0) \neq h\right\}\right] \\
&= P\left(D_i(1) = h, D_i(0) \neq h\right),
\end{aligned}$$

where the second-to-last equality again used the fact that $D_i(1) \neq D_i(0)$ implies $D_i(1) = h$. Similarly,

$$\begin{aligned}
\frac{\partial P\left(D_i = c\right)}{\partial \delta} &= E\left[1\left\{D_i(1) = c\right\} - 1\left\{D_i(0) = c\right\}\right] \\
&= -E\left[1\left\{D_i(1) = h, D_i(0) = c\right\}\right] \\
&= -P\left(D_i(1) = h, D_i(0) = c\right).
\end{aligned}$$

Plugging these expressions into $\partial C/\partial \delta$ yields

$$\begin{aligned}
\frac{\partial C}{\partial \delta} &= \phi_h P\left(D_i(1) = h, D_i(0) \neq h\right) - \phi_c P\left(D_i(1) = h, D_i(0) = c\right) \\
&\quad -\tau pLATE_h P\left(D_i(1) = h, D_i(0) \neq h\right) \\
&= \left(\phi_h - \phi_c S_c - \tau pLATE_h\right) P\left(D_i(1) = h, D_i(0) \neq h\right),
\end{aligned}$$

which is equation (7).

The marginal value of public funds associated with a change in δ is the ratio of the impact on B to the impact on C:

$$MVPF_\delta \equiv \frac{\partial B/\partial \delta}{\partial C/\partial \delta}.$$

By plugging in expressions for these derivatives we obtain

$$MVPF_\delta = \frac{(1-\tau)pLATE_h}{\phi_h - \phi_c S_c - \tau pLATE_h},$$

which is equation (8).

C.2 Rationed Substitutes

We next consider the case where seats in competing programs are rationed. As in Head Start, we assume that seats in the competing program are distributed randomly. Let Z_{ih} and Z_{ic} denote offers in options h and c, and let δ_h and δ_c denote the corresponding offer probabilities. Preferences now depend on both offers. Utilities are described by

$$U_i(h, Z_{ih}),\ U_i(c, Z_{ic}),\ U_i(n),$$

and preschool enrollment choices are defined by

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

Let $\pi_d(z_h, z_c) = P(D_i(z_h, z_c) = d)$ denote the probability of enrollment in option d as a function of the two offers. Total enrollment in option c is

$$P(D_i = c) = \delta_h\delta_c\pi_c(1,1) + \delta_h(1-\delta_c)\pi_c(1,0) + (1-\delta_h)\delta_c\pi_c(0,1) + (1-\delta_h)(1-\delta_c)\pi_c(0,0). \quad (18)$$

We assume that competing preschools adjust δ_c so that $dP(D_i = c)/d\delta_h = 0$. Totally differentiating equation (18) with respect to δ_h yields

$$\frac{d\delta_c}{d\delta_h} = -\frac{\delta_c\left(\pi_c(1,1) - \pi_c(0,1)\right) + (1-\delta_c)\left(\pi_c(1,0) - \pi_c(0,0)\right)}{\delta_h\left(\pi_c(1,1) - \pi_c(1,0)\right) + (1-\delta_h)\left(\pi_c(0,1) - \pi_c(0,0)\right)}.$$

$$= \frac{P\left(D_i(1, Z_{ic}) = h, D_i(0, Z_{ic}) = c\right)}{P\left(D_i(Z_{ih}, 1) = c, D_i(Z_{ih}, 0) \neq c\right)}.$$

To keep enrollment constant, δ_c adjusts by the ratio of the effect of an offer at h on attendance at c to the effect of an offer at c on attendance at c.

Average test scores are given by

$$E\left[Y_i\right] = \delta_h\left(\delta_c E\left[Y_i(D_i(1,1))\right] + (1-\delta_c)E\left[Y_i(D_i(1,0))\right]\right)$$

$$+(1-\delta_h)\left(\delta_c E\left[Y_i(D_i(0,1))\right] + (1-\delta_c)E\left[Y_i(D_i(0,0))\right]\right),$$

so

$$\frac{dE\left[Y_i\right]}{d\delta_h} = \delta_c\left(E\left[Y_i(D_i(1,1)) - Y_i(D_i(0,1))\right]\right)$$

$$+(1-\delta_c)\left(E\left[Y_i(D_i(1,0))\right] - E\left[Y_i(D_i(0,0))\right]\right)$$

$$+\frac{d\delta_c}{d\delta_h}\cdot\left(\delta_h E\left[Y_i(D_i(1,1)) - Y_i(D_i(1,0))\right] + (1-\delta_h)E\left[Y_i(D_i(0,1)) - Y_i(D_i(0,0))\right]\right),$$

which can be rewritten

$$\frac{dE\left[Y_i\right]}{d\delta_h} = E\left[Y_i(D_i(1, Z_{ic})) - Y_i(D_i(0, Z_{ic}))\right]$$

$$+\frac{d\delta_c}{d\delta_h} \cdot (E\left[Y_i(D_i(Z_{ih},1)) - Y_i(D_i(Z_{ih},0))\right])$$

$$= LATE_h \cdot P\left(D_i(1,Z_{ic}) = h, D_i(0,Z_{ic}) \neq h\right)$$

$$+LATE_c \cdot P\left(D_i(1,Z_{ic}) = h, D_i(0,Z_{ic}) = c\right).$$

Here the local average treatment effects are defined as

$$LATE_h = E\left[Y_i(h) - Y_i(D_i(0,Z_{ic}))|D_i(1,Z_{ic}) = h, D_i(0,Z_{ic}) \neq h\right],$$

$$LATE_c = E\left[Y_i(c) - Y_i(D_i(Z_{ih},0))|D_i(Z_{ih},1) = c, D_i(Z_{ih},0) \neq c\right].$$

This can be further simplified to

$$\frac{dE\left[Y_i\right]}{d\delta_h} = (LATE_h + S_c LATE_c) \cdot P\left(D_i(1,Z_{ic}) = h, D_i(0,Z_{ic}) \neq h\right).$$

The effect of an increase in δ_h on the government's budget is

$$\frac{dC}{d\delta_h} = \phi_h \cdot \frac{dP\left(D_i = h\right)}{d\delta_h} - \tau p \cdot \frac{dE\left[Y_i\right]}{d\delta_h}.$$

Since δ_c adjusts to keep $P(D_i = c)$ constant, we have $dP(D_i = c)/d\delta_h = 0$. We assume that all marginal children drawn into c by offers come from n rather than h. This implies $LATE_c = LATE_{nc}$, and furthermore

$$\frac{dP(D_i = h)}{d\delta_h} = P\left(D_i(1,Z_{ic}) = h, D_i(0,Z_{ic}) \neq h\right).$$

Then the marginal value of public funds is

$$MVPF_{\delta,rat} = \frac{dB/d\delta_h}{dC/d\delta_h}$$

$$= (1-\tau)p\left(LATE_h + S_c LATE_{nc}\right) P\left(D_i(1,Z_{ic}) = h, D_i(0,Z_{ic}) \neq h\right)$$

$$\times \left[\phi_h P\left(D_i(1,Z_{ic}) = h, D_i(0,Z_{ic}) \neq h\right) - \tau p\left(LATE_h + S_c LATE_{nc}\right) P\left(D_i(1,Z_{ic}) = h, D_i(0,Z_{ic}) \neq h\right)\right]^{-1}$$

$$= \frac{(1-\tau)p\left(LATE_h + LATE_{nc} \cdot S_c\right)}{\phi_h - \tau p\left(LATE_h + LATE_{nc} \cdot S_c\right)},$$

which is equation (9).

This implies that $MVPF_{\delta,rat} > MVPF_\delta$ whenever Head Start and other preschools have similar test score effects and other preschools are cheapt\er. Specifically, when $LATE_{nc} = LATE_{nh} = LATE > 0$ and $LATE_{ch} = 0$, we have $MVPF_{\delta,rat} = \frac{(1-\tau)pLATE}{\phi_h - \tau pLATE} > MVPF_\delta = \frac{(1-\tau)pLATE}{\frac{\phi_h - \phi_c S_c}{1-S_c} - \tau pLATE}$ whenever $\phi_c < \phi_h$.

C.3 Structural Reforms

Next, consider structural reforms that alter the program feature f. A change in f generates the following impacts on income and the government budget:

$$\frac{\partial B}{\partial f} = (1-\tau)p\frac{\partial E\left[Y_i\right]}{\partial f},$$

$$\frac{\partial C}{\partial f} = \phi_h \frac{\partial P(D_i = h)}{\partial f} + \phi_h'(f)P(D_i = h) + \phi_c \frac{\partial P(D_i = c)}{\partial f} - \tau p \frac{\partial E\left[Y_i\right]}{\partial f}$$

$$= \frac{\partial P\left(D_i = h\right)}{\partial f}\left[\phi_h + \phi_h'(f)\partial\left(\ln P(D_i = h)/\partial f\right)^{-1} + \phi_c \frac{\partial P(D_i = c)/\partial f}{\partial P(D_i = h/\partial f)} - \tau p \frac{\partial E\left[Y_i\right]/\partial f}{\partial P(D_i = h)/\partial f}\right].$$

We can write mean test scores as

$$E\left[Y_i\right] = E\left[Y_i(h)\cdot 1\left\{U_i(h,Z_i) + f \geq U_i(c),\ U_i(h,Z_i) + f \geq 0\right\}\right]$$

$$+E\left[Y_i\left(c\right)\cdot 1\left\{U_i(c) \geq U_i(h,Z_i) + f,\ U_i(c) \geq 0\right\}\right]$$

$$+E\left[Y_i(n)\cdot 1\left\{U_i(h,Z_i) + f \leq 0,\ U_i(c) \leq 0\right\}\right],$$

where we have normalized $U_i(n)$ to zero. The third term in this expression is

$$E\left[Y_i(n)\cdot 1\left\{U_i(h,Z_i) + f \leq 0,\ U_i(c) \leq 0\right\}\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{-f} y\cdot g_{yu}(y,u_h,u_c)du_h du_c dy,$$

where $g_{yu}(\cdot)$ is the joint density function of $Y_i(n)$, $U_i(h,Z_i)$ and $U_i(c)$. Using Leibniz's rule for differentiation under the integral sign and Fubini's theorem, we have

$$\frac{\partial E\left[Y_i(n)\cdot 1\left\{U_i(h,Z_i) + f \leq 0,\ U_i(c) \leq 0\right\}\right]}{\partial f} = \int_{-\infty}^{\infty}\int_{-\infty}^{0}\frac{\partial}{\partial f}\left[\int_{-\infty}^{-f} y\cdot g_{yu}(y,u_h,u_c)du_h\right]du_c dy$$

$$\begin{aligned}
&= -\int_{-\infty}^{\infty}\int_{-\infty}^{0} y\cdot g_{yu}(y,-f,u_c)du_c dy \\
&= -\int_{-\infty}^{0}\left[\int_{-\infty}^{\infty} y\cdot g_{y|u}\left(y|-f,u_c\right)dy\right]g_u(-f,u_c)du_c \\
&= -\int_{-\infty}^{0} E\left[Y_i(n)|U_i(h,Z_i) + f = 0, U_i(c) = u_c\right]g_u(-f,u_c)du_c \\
&= -\int_{-\infty}^{0} g_u(-f,u_c)du_c \cdot E\left[Y_i(n)|U_i(h,Z_i) + f = 0, U_i(c) < 0\right] \\
&= -g_{u_h}(-f)P\left(U_i(c) < 0|U_i(h,Z_i) + f = 0\right)\cdot E\left[Y_i(n)|U_i(h) + f = 0, U_i(c) < 0\right]
\end{aligned}$$

where $g_{y|u}(\cdot)$ is the density of $Y_i(n)$ conditional on the utilities, $g_u(\cdot)$ is the joint density of the utilities, and $g_{u_h}(\cdot)$ is the marginal density of $U_i(h,Z_i)$. The last factor in this expression is the average of $Y_i(n)$ for individuals who are indifferent between Head Start and home care, and strictly

prefer home care to the competing program. The first two factors give the total density associated with this event.

Similar arguments show the effects of a change in f on scores in c and h:

$$\frac{\partial E\left[Y_i\left(c\right)\cdot 1\left\{U_i(c)\geq U_i(h,Z_i)+f,\ U_i(c)\geq 0\right\}\right]}{\partial f} = -g_{c-h}(f)P\left(U_i(c)>0|U_i(h,Z_i)+f=U_i(c)\right) \times E\left[Y_i(c)|U_i(h,Z_i)+f=U_i(c),\ U_i(c)>0\right],$$

$$\frac{\partial E\left[Y_i(h)\cdot 1\left\{U_i(h,Z_i)+f\geq U_i(c),\ U_i(h)+f\geq 0\right\}\right]}{\partial f} = \{g_{c-h}(f)P\left(U_i(c)>0|U_i(h,Z_i)+f=U_i(c)\right) + g_{u_h}(-f)P\left(U_i(c)<0|U_i(h,Z_i)+f=0\right)\} \times E\left[Y_i(h)|U_i(h,Z_i)+f=\max\left\{U_i(c),U_i(n)\right\}\right],$$

where $g_{c-h}(\cdot)$ is the density of $U_i(c)-U_i(h,Z_i)$.

The corresponding effects on choice probabilities are

$$\frac{\partial P\left(D_i=h\right)}{\partial f} = g_{u_h}(-f)P\left(U_i(c)<0|U_i(h,Z_i)+f=0\right) + g_{c-h}(f)P\left(U_i(c)>0|U_i(h,Z_i)+f=U_i(c)\right),$$

$$\frac{\partial P\left(D_i=c\right)}{\partial f} = -g_{c-h}(f)P\left(U_i(c)>0|U_i(h,Z_i)+f=U_i(c)\right).$$

The share of marginal children drawn from the competing program is then given by

$$\overrightarrow{S}_c = -\frac{\partial P(D_i=c)/\partial f}{\partial P(D_i=h)/\partial f}$$

$$= \frac{g_{c-h}(f)P\left(U_i(c)>0|U_i(h,Z_i)+f=U_i(c)\right)}{g_{u_h}(-f)P\left(U_i(c)<0|U_i(h,Z_i)+f=0\right)+g_{c-h}(f)P\left(U_i(c)>0|U_i(h,Z_i)+f=U_i(c)\right)}.$$

By plugging these equations into the expressions for costs and benefits and dividing by the total density of marginal compliers, we obtain

$$MVPF_f = \frac{(1-\tau)pMTE_h}{\phi_h(1+\eta)-\phi_c\overrightarrow{S}_c-\tau pMTE_h},$$

which is equation (12).

C.4 Valuing test score impacts

Here we consider more carefully how to value test score impacts in dollar terms. Specifically, we show that if test score impacts yield corresponding labor supply responses, an adjustment to lifetime earnings impacts is necessary to properly capture the welfare benefits of a policy change.

This argument implies that we should use projected impacts on wages (as opposed to earnings) to value test score gains.

Letting y denote a child's human capital level (as proxied by test scores), we are interested in deriving a child's willingness to pay (as an adult) for an intervention shifting her human capital level from y_0 to $y_1 > y_0$. If this willingness to pay exceeds the net cost to government of financing the human capital increase, then the intervention is efficiency improving in the Kaldor-Hicks sense that all parties *could* be made better off.

We work with a simple static model where children face a competitive labor market with no uncertainty and are free to choose lifetime labor supply in accord with utility maximization. Suppose children have utility over consumption (q) and leisure ($\bar{l}$) given by the function $u\left(q, \bar{l}\right)$. The lifetime budget constraint of a child with human capital level y can be written:

$$q = w(y)\left(T - \bar{l}\right) + b,$$

where $w(y) = (1-\tau)py \equiv \omega$ is the after-tax wage, T is a time endowment, and b is unearned income. The uncompensated (Marshallian) labor supply function is $l(\omega, b)$.

Define the excess expenditure function:

$$e(\omega, \bar{u}) \equiv \min\left\{q - \omega\left(T - \bar{l}\right) : u\left(q, \bar{l}\right) \geq \bar{u}\right\}$$

as the minimal level level of unearned income necessary to obtain utility level $\bar{u}$ at wage level ω. By the envelope theorem

$$\frac{\partial}{\partial \omega} e(\omega, \bar{u}) = -l_c(\omega, \bar{u}),$$

where $l_c(\omega, \bar{u})$ is the compensated (Hicksian) labor supply function.

Suppose that at human capital level y_0 the child is able to obtain utility level u_0. The compensating variation:

$$CV(y_0, y_1) \equiv e(w(y_0), u_0) - e(w(y_1), u_0),$$

measures how much income a child could give away at human capital level y_1 and still obtain his old utility level u_0. A first order Taylor approximation yields:

$$\begin{aligned} CV(y_0, y_1) &\approx (1-\tau) p l_c(w(y_0), u_0)(y_1 - y_0) \\ &= (1-\tau) p l(w(y_0), b)(y_1 - y_0). \end{aligned} \tag{19}$$

In words, the value to a child of a small increase in test scores is given by the mechanical impact this increase in her wage would have on her lifetime earnings if her labor supply were fixed at $l(w(y_0), b)$.

This is to be contrasted with the actual effect of the human capital increase on his earnings

which can be written:

$$w\left(y_1\right) l\left(w\left(y_1\right), b\right) - w\left(y_0\right) l\left(w\left(y_0\right), b\right) \approx \left(1-\tau\right) pl\left(w\left(y_0\right), b\right)\left(1+\epsilon\right)\left(y_1 - y_0\right)$$

where $\epsilon \equiv \frac{w(y_0)}{l(w(y_0),b)} \frac{\partial}{\partial \omega} l\left(w\left(y_0\right), b\right)$ gives the uncompensated elasticity of labor supply. Relative to (19), this expression has an extra term $(1+\epsilon)$ that reflects how the child adjusts her lifetime labor supply in response to the increase in her after-tax wage. By the envelope theorem, these behavioral changes (when they are small) do not yield additional utility.

The upshot of this analysis is that empirical estimates of the impact of test scores on earnings need to be deflated by $\frac{1}{1+\epsilon}$ to reflect the child's valuation of the intervention. Much of the literature finds small (or even negative) long run uncompensated labor supply elasticities suggesting that the necessary adjustment is probably small (Ashenfelter et al., 2010; Blundell et al., 2015). Consistent with this view, Lindqvist and Vestman (2011) find the proportional response of wages to test scores to be only slightly below the corresponding response of earnings (see Appendix Table A.IV).

Appendix D: Empirical Cost Benefit Analysis

This appendix discusses in more detail the assumptions underlying the cost-benefit analysis of Section VI.

Representativeness of the HSIS data

The HSIS data are a nationally-representative random sample of Head Start applicants, and HSIS offers are distributed randomly (Puma et al., 2010). The HSIS is therefore ideal for estimating values of $LATE_h$ and S_c *in the population of Head Start applicants.*[21] Fortunately, the current Head Start application rate is high, which limits the scope for selection into the applicant pool that might change with program scale. Currie (2006) reports that two-thirds of eligible children participated in Head Start in 2000. This is higher than the Head Start participation rate in the HSIS sample (49 percent). However, fifteen percent of participants attend undersubscribed centers outside the HSIS sample, which implies that about 57 percent $(0.85 \cdot 0.49 + 0.15)$ of all applicants participate in Head Start (Puma et al., 2010). For this to be consistent with a participation rate of two-thirds among eligible households, virtually all eligible households must apply. Therefore, selection into the Head Start applicant pool is unlikely to be quantitatively important for our analysis.

Program benefits

The term p in equation (4) gives the dollar value of a one standard deviation increase in test scores. Although earnings are unavailable for the HSIS sample, a growing body of evidence shows a consistent link between short-run test score effects and earnings impacts. Rather than choose a particular value for p, we consider a range of values consistent with the literature, focusing on how low of a value would be necessary to undermine the conclusion that Head Start pays for itself.

Appendix Table A.IV summarizes several studies that compare test score and earnings impacts for the same intervention. The most closely related study is by Chetty et al. (2011), an analysis of the Tennessee STAR class size experiment. Chetty et al. (2011, p.7 online appendix) show that a one standard deviation increase in kindergarten test scores induced by an experimental change in classroom quality yields a 13.1 percent increase in earnings at age 27.[22] The STAR results also

[21] As detailed in Appendix A, our analysis excludes HSIS applicants without followup data (20 percent of the sample), and we use weights that capture the probability a child is assigned to Head Start but not the probability a Head Start center is sampled from the larger population of centers. Our estimates may not be representative of the full population of Head Start applicants if children without followup data differ systematically from other children or if applicant populations differ in a way that is systematically related to center-level sampling probabilities.

[22] Effects in standard deviation units may have different meanings if score distributions differ across populations or over time. For example, Cascio and Staiger (2012) show that test score norming partially explains fadeout in effects of educational interventions. Sojourner (2009) shows that the standard deviation of nationally-normed scores in the STAR sample is 87 percent of the national standard deviation. The standard deviations of Spring 2003 PPVT and WJIII scores in the HSIS are 70 percent and 91 percent of the national standard deviation, for a mean of 81 percent. This suggests we should rescale the STAR estimate of 13.1 percent to 12.2 percent in our sample; our baseline calibrations use a more conservative estimate of 10 percent.

suggest that immediate test score effects of early-childhood programs predict earnings gains better than test score effects in other periods: classrooms that boost test scores in the short run increase earnings in the long run despite fadeout of test score impacts in the interim. We therefore project earnings gains based on our first-year estimates of $LATE_h$.

The STAR classroom quality estimate of 13.1 percent is smaller than a corresponding OLS estimate controlling for rich family characteristics in the STAR sample (18 percent), and comparable to estimates from Chetty et al. (2014b) linking test score and earnings impacts for teacher value-added (10.3 percent for value-added, 12 percent for OLS with controls). The Chetty et al. (2014b) findings also replicate the pattern of long-run earnings impacts coupled with fadeout of medium-run test score effects. In an analysis of the Perry Preschool Project, Heckman et al. (2010b) estimate larger ratios of earnings per standard deviation of test scores (24 to 29 percent). Sibling fixed effects estimates from studies of Head Start by Currie and Thomas (1995) and Garces et al. (2002) suggest much larger ratios, though the earnings estimates are also very statistically imprecise. To be conservative, our baseline calibrations assume an earnings impact of 10 percent per standard deviation of earnings, which is at the bottom of the range of estimates reported in Table A.IV.[23]

Calculating percentage changes in earnings requires a prediction of average earnings in the HSIS population. Chetty et al. (2011) calculate that the average present discounted value of earnings in the United States is approximately $522,000 at age 12 in 2010 dollars. Using a 3-percent discount rate, this yields a present discounted value of $438,000 at age 3.4 (the average age of applicants in the HSIS). Children who participate in Head Start are disadvantaged and therefore likely to earn less than the US average. The average household participating in Head Start earned 46 percent of the US average in 2013 (US DHHS, 2013; Noss, 2014). Lee and Solon (2009) find an average intergenerational income elasticity in the United States of roughly 0.4, implying that the average child in Head Start is expected to earn 78 percent of the US average $(1 - (1 - 0.46) \cdot 0.4)$.[24] These calculations yield a present value of earnings $\bar{e}$ equal to $\$343,492$ at age 3.4.

Thus, our baseline estimate is that the marginal benefit of enrolling an additional child in Head Start is $0.1 \cdot \$343,492 \cdot LATE_h$. Using the pooled first-year estimate of $LATE_h$ reported in Section III, we project an earnings impact of $0.1 \cdot \$343,492 \cdot 0.247 = \$8,472$. We set $\tau = 0.35$ based upon estimates from the Congressional Budget Office (2012, Figure 2) that account for federal and state taxes along with food stamps participation. This generates a discounted after-tax lifetime earnings gain of $5,513 for compliers.

[23]The only estimates below 10 percent in Table A.IV are from Murnane et al. (1995) and Currie and Thomas (1999). Murnane et al. use High School and Beyond data to construct an OLS estimate relating 12th grade scores to log wages at age 24 for males (7.7 percent). The same approach produces a larger estimate for females (10.9 percent). Currie and Thomas report partial effects from models that include both math and reading scores. Since these scores are very highly correlated, the total effect for a single test score is likely to be larger.

[24]Chetty et al. (2014c) find that the IGE is not constant across the parent income distribution. Appendix Figure IA in their study shows that the elasticity of mean child income with respect to mean parent income is 0.414 for families between the 10th and 90th percentile of parent income but lower for families below the 10th percentile. Since Head Start families are drawn from these poorer populations, it is reasonable to expect that the relevant IGE for this population is below 0.4, implying that our rate of return calculations are conservative.

Program costs

Equation (7) shows that the net marginal social cost of Head Start enrollment depends on the costs to government of enrollment in Head Start and competing preschools along with the share of compliers drawn from other preschools. Per-pupil expenditure in Head Start is approximately $\$8,000$ (US DHHS, 2013). As reported in Column (7) of Table III, the estimated share of compliers drawn from other preschools is 0.34.

To get an idea of the costs of competing programs, Panel A of Appendix Table A.II reports information on funding sources for Head Start and competing preschool centers. These data come from a survey administered to the directors of Head Start centers and other centers attended by children in the HSIS experiment. Column (2) shows that competing preschools receive financing from a mix of sources, and many receive public subsidies. Thirty-nine percent of competing centers did not complete the survey, but among respondents, only 25 percent (0.153/0.606) report parent fees as their largest source of funding. The modal funding source is state preschool programs (30 percent), and an additional 16 percent report that other childcare subsidies are their primary funding source. Column (3) reports characteristics of competing preschools attended by c-compliers, estimated using a generalization of the methods for characterizing compliers described by Abadie (2002) (see Appendix B). In the absence of a Head Start offer, c-compliers attend preschools that rely slightly more on parent fees, but most are financed by a mix of state preschool programs, childcare subsidies, and other funding sources.

Panel B of Table A.II compares key inputs and practices in Head Start and competing preschool centers attended by children in the HSIS sample. On some dimensions, Head Start centers appear to provide higher-quality services than competing programs. Columns (4) and (5) show that Head Start centers are more likely to provide transportation to preschool and frequent home visiting than competing centers. Average class size is also smaller in Head Start, and Head Start center directors have more experience than their counterparts in competing preschools. As a result of these differences, Head Start centers score higher on a composite measure of quality. On the other hand, teachers at alternative programs are more likely to have bachelors degrees and certification, and these programs are more likely to provide full-day service. Column (6) shows that competing preschools attended by Head Start compliers are very similar to the larger set of alternative preschools in the HSIS sample.

Table A.II suggests that roughly 75% of competing programs are financed *primarily* by public subsidies. Of course, even centers that are financed primarily by fees are likely to receive subsidies for enrolling the disadvantaged students in our sample (who are unlikely to be able to pay full price). Based upon this, we use as our "preferred" estimate that $\phi_c = 0.75\phi_h$, which is a conservative estimate if Head Start and competing preschools are equally costly and 75% of Head Start eligible students had their tuition fully subsidized at competing preschools while others receive partial subsidies. Our "pessimistic" scenario where $\phi_c = 0.5\phi_h$ corresponds roughly to the case where all of the non-responding centers in Table A.II relied on private fees for financing. Finally, the "naive" assumption that $\phi_c = 0$ is useful as a benchmark for assessing the importance of fiscal externalities.

Appendix E: Interacted Two-stage Least Squares

This Appendix investigates the use of the interacted two-stage least squares approach described in Section VII to estimate models treating both Head Start and other preschools as endogenous variables. We begin with a simple example that clarifies the parameters estimated by this strategy, then apply the strategy to the HSIS data.

Interacted 2SLS estimand

Suppose there is a single binary covariate $X_i \in \{0, 1\}$. Under the assumptions described in Section IV, covariate-specific instrumental variables coefficients give local average treatment effects:

$$\frac{E\left[Y_i | Z_i = 1, X_i = x\right] - E\left[Y_i | Z_i = 0, X_i = x\right]}{E\left[1\left\{D_i = h\right\} | Z_i = 1, X_i = x\right] - E\left[1\left\{D_i = h\right\} | Z_i = 0, X_i = x\right]} = LATE_h(x).$$

Furthermore, we have

$$LATE_h(x) = S_c(x) LATE_{ch}(x) + (1 - S_c(x)) LATE_{nh}(x),$$

where $S_c(x) = \frac{P(D_i(1)=h, D_i(0)=c | X_i=x)}{P(D_i(1)=h, D_i(0)\neq h | X_i=x)}$ is the covariate-specific share of compliers drawn from other preschools. The $S_c(x)$ are identified, but if we assume $LATE_{ch}$ and $LATE_{nh}$ vary with x in an unrestricted way we have two equations in four unknowns and cannot use the available information to recover subLATEs.

Suppose instead we assume that the subLATEs don't vary with x, so that $LATE_{dh}(x) = LATE_{dh}\ \forall x,\ d \in \{c, n\}$. Our two equations are

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

$$LATE_h(0) = S_c(0) LATE_{ch} + (1 - S_c(0)) LATE_{nh}.$$

The solution to this system is

$$LATE_{nh} = \frac{S_c(0) LATE_h(1) - S_c(1) LATE_h(0)}{S_c(0) - S_c(1)},$$

$$LATE_{ch} = \frac{(1 - S_c(0)) LATE_h(1) - (1 - S_c(1)) LATE_h(0)}{(1 - S_c(0)) - (1 - S_c(1))}.$$

The right-hand sides tell us the probability limits of 2SLS coefficients from a model instrumenting $1\left\{D_i = h\right\}$ and $1\left\{D_i = c\right\}$ with Z_i and $Z_i \cdot X_i$ and controlling for X_i. Specifically, the Head Start coefficient from this interacted 2SLS strategy equals $LATE_{nh}$ and the other preschool coefficient equals $LATE_{nh} - LATE_{ch}$. To see this note that the 2SLS system is just-identified under constant effects which implies constant subLATEs. There is therefore exactly one way to solve for the two effects of interest using the available information; since the equations above yield these effects they must give this solution.

If the constant effects assumption is wrong, the interacted 2SLS strategy yields a Head Start coefficient equal to

$$LATE_{nh} = \frac{S_c(0)S_c(1)}{S_c(0)-S_c(1)} LATE_{ch}(1) + \frac{S_c(0)(1-S_c(1))}{S_c(0)-S_c(1)} LATE_{nh}(1)$$

$$- \frac{S_c(1)S_c(0)}{S_c(0)-S_c(1)} LATE_{ch}(0) - \frac{S_c(1)(1-S_c(0))}{S_c(0)-S_c(1)} LATE_{nh}(0),$$

which can be written

$$\begin{aligned} LATE_{nh} = & \frac{S_c(0)S_c(1)}{S_c(0)-S_c(1)} \cdot (LATE_{ch}(1) - LATE_{ch}(0)) \\ & + (w_n(1) LATE_{nh}(1) + (1 - w_n(1)) LATE_{nh}(0)) , \end{aligned} \tag{20}$$

where

$$w_n(1) = \frac{S_c(0)(1 - S_c(1))}{S_c(0) - S_c(1)}.$$

This expression shows that the interacted 2SLS strategy yields a Head Start coefficient equal to a weighted average of the subLATEs $LATE_{nh}(x)$, plus a term that depends on heterogeneity in $LATE_{ch}(x)$. If there is heterogeneity in this other subLATE, this strategy does not recover the causal effect of h relative to n for any well-defined subpopulation. This result is a special case of the results in Kirkboen et al. (2014) and Hull (2015), who show that 2SLS does not generally recover causal effects in models with multiple endogenous variables.

Appendix F: Selection Model

F.1 Control Functions

This appendix derives the control function terms for the selection model of Section VII. Households participate in Head Start ($D_i = h$) when

$$\psi_h(X_i, Z_i) + v_{ih} > \psi_c(X_i) + v_{ic},\ \psi_h(X_i, Z_i) + v_{ih} > 0\ ,$$

which can be re-written

$$\frac{v_{ic} - v_{ih}}{\sqrt{2\left(1 - \rho(X_i)\right)}} < \frac{\psi_h(X_i, Z_i) - \psi_c(X_i)}{\sqrt{2(1 - \rho(X_i))}},\ -v_{ih} < \psi_h(X_i).$$

The random variables $\left(\frac{v_{ic}-v_{ih}}{\sqrt{2(1-\rho(X_i)}}\right)$ and $(-v_{ih})$ have a bivariate standard normal distribution with correlation $\sqrt{\frac{1-\rho(X_i)}{2}}$. Then using the formulas in Tallis (1961) for the expectations of bivariate standard normal random variables truncated from above, we have

$$\begin{aligned} E\left[\frac{v_{ic} - v_{ih}}{\sqrt{2\left(1 - \rho(X_i)\right)}} | X_i, Z_i, D_i = h\right] = &\ \Lambda\left(\tfrac{\psi_h(X_i,Z_i)-\psi_c(X_i)}{\sqrt{2(1-\rho(X_i))}}, \psi_h(X_i); \sqrt{\tfrac{1-\rho(X_i)}{2}}\right), \\ E\left[-v_{ih} | X_i, Z_i, D_i = h\right] = &\ \Lambda\left(\psi_h(X_i), \tfrac{\psi_h(X_i,Z_i)-\psi_c(X_i)}{\sqrt{2(1-\rho(X_i))}}; \sqrt{\tfrac{1-\rho(X_i)}{2}}\right), \end{aligned}$$

where

$$\Lambda(a_1, b_1; \xi) \equiv -\left[\frac{\phi(a_1)\Phi\left(\frac{b_1-\xi a_1}{\sqrt{1-\xi^2}}\right) + \xi\phi\left(b_1\right)\Phi\left(\frac{a_1-\xi b_1}{\sqrt{1-\xi^2}}\right)}{\Phi_b(a_1, b_1; \xi)}\right].$$

Here $\phi(\cdot)$ and $\Phi(\cdot)$ are the PDF and CDF of the standard normal distribution, while $\Phi_b(\cdot)$ is the bivariate standard normal CDF.

Defining $\lambda_d\left(X_i, Z_i, D_i\right) \equiv E\left[v_{id} | X_i, Z_i, D_i\right]$, this implies that we can write

$$\begin{aligned} \lambda_h(X_i, Z_i, h) = &\ -\Lambda\left(\psi_h(X_i), \tfrac{\psi_h(X_i,Z_i)-\psi_c(X_i)}{\sqrt{2(1-\rho(X_i))}}; \sqrt{\tfrac{1-\rho(X_i)}{2}}\right), \\ \lambda_c(X_i, Z_i, h) = &\ -\Lambda\left(\psi_h(X_i), \tfrac{\psi_h(X_i,Z_i)-\psi_c(X_i)}{\sqrt{2(1-\rho(X_i))}}; \sqrt{\tfrac{1-\rho(X_i)}{2}}\right) \\ &+ \sqrt{2(1-\rho(X_i))} \cdot \Lambda\left(\tfrac{\psi_h(X_i,Z_i)-\psi_c(X_i)}{\sqrt{2(1-\rho(X_i))}}, \psi_h(X_i); \sqrt{\tfrac{1-\rho(X_i)}{2}}\right). \end{aligned}$$

Similar calculations for $D_i = c$ and $D_i = n$ yield

$$\begin{aligned}
\lambda_h(X_i, Z_i, c) = & -\Lambda\left(\psi_c(X_i), \tfrac{\psi_c(X_i)-\psi_h(X_i,Z_i)}{\sqrt{2(1-\rho(X_i))}}; \sqrt{\tfrac{1-\rho(X_i)}{2}}\right) \\
& + \sqrt{2\left(1-\rho(X_i)\right)} \cdot \Lambda\left(\tfrac{\psi_c(X_i)-\psi_h(X_i,Z_i)}{\sqrt{2(1-\rho(X_i))}}, \psi_c(X_i); \sqrt{\tfrac{1-\rho(X_i)}{2}}\right), \\
\lambda_c(X_i, Z_i, c) = & -\Lambda\left(\psi_c(X_i), \tfrac{\psi_c(X_i)-\psi_h(X_i,Z_i)}{\sqrt{2(1-\rho(X_i))}}; \sqrt{\tfrac{1-\rho(X_i)}{2}}\right), \\
\lambda_h(X_i, Z_i, n) = & \Lambda\left(-\psi_h(X_i, Z_i), -\psi_c(X_i); \rho(X_i)\right), \\
\lambda_c(X_i, Z_{i,}, n) = & \Lambda\left(-\psi_c(X_i), -\psi_h(X_i, Z_i); \rho(X_i)\right).
\end{aligned}$$

F.2 Identification

We next consider identification of the selection model parameters and the subLATEs in a model with one binary covariate, $X_i \in \{0, 1\}$. In this case the choice model is fully saturated and there are four parameters for each value of X_i: $\psi_h(x, 1)$, $\psi_h(x, 0)$, $\psi_c(x)$, and $\rho(x)$. These parameters are just-identified and perfectly fit the four independent conditional choice probabilities

$$\pi_d(x, z) = Pr\left[D_i = d | X_i = x, Z_i = z\right], \; d \in \{h, c\}, \; z \in \{0, 1\}.$$

The parameters of the selection model are therefore implicit functions of the choice probabilities.

Let $\Delta_d(x)$ denote the difference in mean outcomes between offered and non-offered children, conditional on X_i and D_i:

$$\Delta_d(x) = E\left[Y_i | X_i = x, Z_i = 1, D_i = d\right] - E\left[Y_i | X_i = x, Z_i = 0, D_i = d\right].$$

Evaluating equation (16) for $X_i = 1$ and $X_i = 0$ gives

$$\Delta_d(1) = \gamma_{dh}\left(\lambda_h(1, 1, d) - \lambda_h(1, 0, d)\right) + \gamma_{dc}\left(\lambda_c(1, 1, d) - \lambda_c(1, 0, d)\right),$$

$$\Delta_d(0) = \gamma_{dh}\left(\lambda_h(0, 1, d) - \lambda_h(0, 0, d)\right) + \gamma_{dc}\left(\lambda_c(0, 1, d) - \lambda_c(0, 0, d)\right).$$

Solving these equations for the selection coefficients yields

$$\gamma_{dh} = \frac{\Delta_d(1)\left(\lambda_c(0, 1, d) - \lambda_c(0, 0, d)\right) - \Delta_d(0)\left(\lambda_c(1, 1, d) - \lambda_c(1, 0, d)\right)}{\left(\lambda_h(1, 1, d) - \lambda_h(1, 0, d)\right)\left(\lambda_c(0, 1, d) - \lambda_c(0, 0, d)\right) - \left(\lambda_h(0, 1, d) - \lambda_h(0, 0, d)\right)\left(\lambda_c(1, 1, d) - \lambda_c(1, 0, d)\right)},$$

$$\gamma_{dc} = \frac{\Delta_d(1)\left(\lambda_h(0, 0, d) - \lambda_h(0, 1, d)\right) - \Delta_d(0)\left(\lambda_h(1, 0, d) - \lambda_h(1, 1, d)\right)}{\left(\lambda_h(1, 1, d) - \lambda_h(1, 0, d)\right)\left(\lambda_c(0, 1, d) - \lambda_c(0, 0, d)\right) - \left(\lambda_h(0, 1, d) - \lambda_h(0, 0, d)\right)\left(\lambda_c(1, 1, d) - \lambda_c(1, 0, d)\right)}.$$

These expressions have the form of multivariate instrumental variables coefficients. Specifically, they are coefficients from an infeasible IV model that uses Z_i and $Z_i X_i$ as instruments for v_{ih} and v_{ic} in the $D_i = d$ sample, controlling for a main effect of X_i. Though v_{ih} and v_{ic} are unobserved,

the $\lambda_d(X_i, Z_i, D_i)$ functions capture their conditional means and can therefore be used to construct the first stage for the IV model.

The expressions for γ_{dh} and γ_{dc} have the same denominator. A necessary and sufficient condition for identification of the two selection coefficients is that this denominator is non-zero. To interpret the requirements for identification, note that the $\lambda_d(\cdot)$ are functions of the selection model parameters, so they are implicitly functions of the choice probabilities $\pi(x, z)$. This implies that if $\pi_d(x, 1) = \pi_d(x, 0)\ \forall d$, then $\lambda_h(x, 1, d) = \lambda_h(x, 0, d)$ and $\lambda_c(x, 1, d) = \lambda_c(x, 0, d)$, resulting in a denominator equal to zero. A necessary condition for identification is therefore that the Head Start offer shifts choice probabilities for both covariate groups. Similarly, if $\pi_d(1, z) = \pi_d(0, z)\ \forall d$ for either $z = 0$ or $z = 1$, the denominator equals zero. A second necessary condition is therefore that choice probabilities differ across covariate groups conditional on the Head Start offer. This requires differences in compliance group shares (always takers, c-never takers, n-never takers, c-compliers and n-compliers). Finally, note that the denominator may be zero even if the offer shifts behavior for both covariate groups and choice probabilities differ conditional on Z_i. Identification requires Head Start offers to shift the conditional means of both v_{ih} and v_{ic} in such a way that the mean changes in the two unobservables are not proportional.

F.3 Estimating SubLATEs

After estimating the selection model we use it to predict mean potential outcomes for subpopulations that respond differently to the Head Start offer. We then use these predictions to compute treatment effects and assess the fit of the model. For example, we construct estimates of $LATE_{nh}$, the effect of Head Start relative to home care for children that switch from home care to Head Start in response to an offer.

N-compliers switch from n to h when offered, and are therefore described by

$$\psi_h(X_i, 1) + v_{ih} > 0 > \psi_h(X_i, 0) + v_{ih},\ \psi_c(X_i) + v_{ic} < 0.$$

We can rewrite these conditions

$$-\psi_h(X_i, 1) < v_{ih} < -\psi_h(X_i, 0),\ v_{ic} < -\psi_c(X_i).$$

The selection errors v_{ih} and v_{ic} are truncated between $(-\psi_h(X_i, 1), -\psi_h(X_i, 0))$ and $(-\infty, -\psi_c(X_i))$ for n-compliers. Equation (14) therefore implies that mean potential outcomes for n-compliers are

$$E\left[Y_i(d)|X_i, -\psi_h(X_i, 1) < v_{ih} < -\psi_h(X_i, 0),\ v_{ic} < -\psi_c(X_i)\right] = \theta_{d0} + X_i'\theta_{dx}$$

$$+\gamma_{dh}\Lambda_0\left(-\psi_h(X_i, 1), -\psi_h(X_i, 0), -\infty, -\psi_c(X_i); \rho(X_i)\right)$$

$$+\gamma_{dc}\Lambda_0\left(-\infty, -\psi_c(X_i), -\psi_h(X_i, 1), -\psi_h(X_i, 0); \rho(X_i)\right),$$

where

$$\Lambda_0(a_0,a_1,b_0,b_1;\xi) = \frac{\phi(a_0)\left[\Phi\left(\frac{b_1-\xi a_0}{\sqrt{1-\xi^2}}\right)-\Phi\left(\frac{(1-\xi)a_0}{\sqrt{1-\xi^2}}\right)\right]-\phi(a_1)\left[\Phi\left(\frac{b_1-\xi a_1}{\sqrt{1-\xi^2}}\right)-\Phi\left(\frac{b_0-\xi a_1}{\sqrt{1-\xi^2}}\right)\right]}{\Phi_b\left(a_1,b_1;\xi\right)-\Phi_b(a_1,b_0;\xi)-\Phi_b(a_0,b_1;\xi)+2\Phi_b(a_0,b_0;\xi)}$$
$$+\frac{\xi\phi(b_0)\left[\Phi\left(\frac{a_1-\xi b_1}{\sqrt{1-\xi^2}}\right)-\Phi\left(\frac{a_0-\xi b_0}{\sqrt{1-\xi^2}}\right)\right]-\xi\phi(b_1)\left[\Phi\left(\frac{a_1-\xi b_1}{\sqrt{1-\xi^2}}\right)-\Phi\left(\frac{a_0-\xi b_1}{\sqrt{1-\xi^2}}\right)\right]}{\Phi_b\left(a_1,b_1;\xi\right)-\Phi_b(a_1,b_0;\xi)-\Phi_b(a_0,b_1;\xi)+2\Phi_b(a_0,b_0;\xi)}.$$

The $\Lambda_0(\cdot)$ function gives means of bivariate standard normal random variables truncated from both sides (Tallis, 1961). Analogous derivations give mean potential outcomes for c-compliers, always takers, n-never takers, and c-never takers.

An estimate of mean $Y_i(d)$ for n compliers with covariates X_i is given by

$$\begin{aligned}\hat{\mu}_d^{nh}(X_i) \quad &= \hat{\theta}_{d0} + X_i'\hat{\theta}_{dx} \quad +\hat{\gamma}_{dh}\Lambda_0\left(-\hat{\psi}_h(X_i,1),-\hat{\psi}_h(X_i,0),-\infty,-\hat{\psi}_c(X_i);\hat{\rho}(X_i)\right)\\ &\qquad\qquad +\hat{\gamma}_{dc}\Lambda_0\left(-\infty,-\hat{\psi}_c(X_i),-\hat{\psi}_h(X_i,1),-\hat{\psi}_h(X_i,0);\hat{\rho}(X_i)\right),\end{aligned}$$

where $\hat{\psi}_h$ and $\hat{\rho}$ come from a first-step multinomial probit model and $\hat{\theta}_d$, $\hat{\theta}_d^x$, $\hat{\gamma}_d^h$ and $\hat{\gamma}_d^c$ come from a second-step least squares regression. To obtain unconditional estimates, we integrate over the distribution of X_i for n-compliers. An estimate of the marginal mean of $Y_i(d)$ for n-compliers is given by

$$\hat{\mu}_d^{nh} = \sum_i \left(\frac{\hat{\omega}_i^{nh}}{\sum_j \hat{\omega}_j^{nh}}\right)\hat{\mu}_d^{nh}(X_i),$$

where

$$\hat{\omega}_i^{nh} = \left[\Phi_b\left(-\hat{\psi}_h(X_i,0),-\hat{\psi}_c(X_i);\hat{\rho}(X_i)\right)-\Phi_b\left(-\hat{\psi}_h(X_i,1),-\hat{\psi}_c(X_i);\hat{\rho}(X_i)\right)\right]w_i$$

is an estimate of the probability that individual i is an n-complier conditional on his or her covariates, multiplied by the HSIS sample weight w_i. We then construct the subLATE estimate $\hat{LATE}_{nh} = \hat{\mu}_h^{nh} - \hat{\mu}_n^{nh}$. Estimates of mean potential outcomes and treatment effects for other subgroups are obtained via similar calculations.

F.4 Specification tests

Testing for underidentification

The identification argument in Section F.2 shows that the selection coefficients for enrollment alternative d are identified when there exist an x and x' in the support of X_i such that

$$\begin{aligned}&\left(\lambda_h(x,1,d)-\lambda_h(x,0,d)\right)\left(\lambda_c(x',1,d)-\lambda_c(x',0,d)\right)\neq\\ &\left(\lambda_h(x',1,d)-\lambda_h(x',0,d)\right)\left(\lambda_c(x,1,d)-\lambda_c(x,0,d)\right).\end{aligned}$$

Equivalently, γ_{dh} and γ_{dc} are not identifed if

$$\lambda_h(x,1,d) - \lambda_h(x,0,d) = q_{d1} \times (\lambda_c(x,1,d) - \lambda_c(x,0,d)) \ \forall x$$

for some proportionality factor q_d. We test the null hypothesis that the model is underidentified by fitting the least squares regression

$$\hat{\lambda}_h(X_i,1,d) - \hat{\lambda}_h(X_i,0,d) = \sum_{k=0}^{3} q_{dk} \left(\hat{\lambda}_c(X_i,1,d) - \hat{\lambda}_c(X_i,0,d)\right)^k + \eta_{id} \tag{21}$$

in the sample with $D_i = d$. The null hypothesis that $q_{d0} = q_{d2} = q_{d3} = 0$ is compatible with underidentification of the outcome equation for alternative d; if this hypothesis is false, the control function differences are not proportional and the selection parameters are identified.

To account for estimation error in the first-step multinomial probit parameters we conduct inference via the nonparametric bootstrap. Let $\hat{q}_d = (\hat{q}_{d0}, \hat{q}_{d2}, \hat{q}_{d3})'$ denote full-sample estimates from equation (21) and let $\hat{q}_d^b$ denote corresponding estimates in bootstrap sample b. We form the test statistic

$$\hat{F}_d = \frac{\hat{q}_d' \hat{V}_{qd}^{-1} \hat{q}_d}{3},$$

where

$$\hat{V}_{qd} = \frac{1}{T} \sum_{b=1}^{T} \left(\hat{q}_d^b - \bar{q}_d\right) \left(\hat{q}_d^b - \bar{q}_d\right)'$$

and $\bar{q}_d$ is the mean of $\hat{q}_d^b$ across bootstrap samples. We then compare $\hat{F}_d$ to critical values of the $F(3,\infty)$ distribution. The results of this test are reported in Appendix Figure A.II.

Testing additive separability

The key restriction in equation (14) is additive separability: mean potential outcomes are additively separable in X_i, v_{ih} and v_{ic}. As a result, the selection coefficients do not depend on X_i and these coefficients can be identified via comparisons of gaps in selected outcomes by offer status across covariate groups. The additive separability restriction cannot be tested with a single binary covariate, but it is testable if X_i takes more than two values.

To test the additive separability restriction for care alternative d we estimate regressions of the form

$$\hat{\epsilon}_{id} = \tilde{\theta}_{d0} + X_i'\tilde{\theta}_{dx} + \tilde{\gamma}_{dh}\hat{\lambda}_h(X_i,Z_i,d) + \tilde{\gamma}_{dc}\hat{\lambda}_c(X_i,Z_i,d) + \hat{\lambda}_h(X_i,Z_i,d)X_i'\xi_{dh} + \hat{\lambda}_c(X_i,Z_i,d)X_i'\xi_{dc} + u_{id}$$

for each care alternative, where $\hat{\epsilon}_{id}$ is the residual from two-step estimation of (15). We then construct an F-statistic for the joint null hypothesis that $\xi_{dh} = \xi_{dc} = 0$ for all three care alternatives. Let $\hat{F}$ denote the full-sample F-statistic for this test, and let $\hat{\xi}_{dh}$ and $\hat{\xi}_{dc}$ denote full-sample estimates of ξ_{dh} and ξ_{dc}. In bootstrap sample b we form corresponding estimates $\hat{\xi}_{dh}^b$ and $\hat{\xi}_{dc}^b$ and test the hypothesis that $\hat{\xi}_{dh}^b = \hat{\xi}_{dh}$ and $\hat{\xi}_{dc}^b = \hat{\xi}_{dc}$ for all d, generating the test statistic $\hat{F}^b$. A bootstrap p-value for a score test of additive separability is then

$$p_T = \frac{1}{T}\sum_{b=1}^{T} 1\left[\hat{F}^b > \hat{F}\right].$$

Table VII reports p-values for this test.

Testing model fit

Our control function approach requires correct specification of both the choice model and the model for outcomes. To assess the fit of the choice model we use the multinomial probit estimates to predict probabilities of Head Start and substitute preschool participation, $\hat{\pi}_h(X_i, Z_i)$ and $\hat{\pi}_c(X_i, Z_i)$. We then split the sample into 25 cells defined by interactions of quintiles of the two probabilities. Cells with fewer than 50 observations are grouped into a single cell. Finally, we test that empirical choice probabilities match mean predicted probabilities in each cell, treating the mean predictions as fixed. Appendix Figure A.I plots empirical choice probabilities against cell means of the two model predictions. The nonparametric means are very close to the model predictions and a joint test of equality does not reject. This suggests that the choice model fits well.

Two additional analyses assess the fit of the model for outcomes. The first splits the sample into vingtiles of predicted $LATE_h$, and compares model-predicted estimates to IV estimates within these bins. As shown in Appendix Figure A.III, the model predictions tightly matches the IV estimates while also capturing substantial effect heterogeneity. We cannot reject that the IV estimates and model predictions are equal up to sampling error ($p = 0.26$).

The second analysis compares instrumental variables estimates of mean potential outcomes that are nonparametrically identified to corresponding estimates from the selection model. As shown in Appendix B, for example, an estimate of mean $Y_i(n)$ for n-compliers can be obtained by estimating the instrumental variables model

$$Y_i 1\{D_i = n\} = \kappa_0 + \kappa_n 1\{D_i = c\} + u_i,$$

$$1\{D_i = n\} = m_0 + m_1 Z_i + e_i.$$

The IV estimate $\hat{\kappa}_n$ is a consistent estimate of $E[Y_i(n)|D_i(1) = h, D_i(0) = n]$, which can be compared to the two-step control function estimate $\hat{\mu}_n^{nh}$.

We use a bootstrap covariance matrix to test the fit of the outcome model. Let $\hat{\tau}$ denote a vector of differences between nonparametrically estimated and model-predicted moments (for example, $\hat{\kappa}_n - \hat{\mu}_n^{nh}$), and let $\hat{\tau}_b$ denote the corresponding estimate in bootstrap sample b. We form the test statistic

$$\hat{W} = \hat{\tau}' \hat{V}_\tau^{-1} \hat{\tau}$$

where

$$\hat{V}_\tau = \frac{1}{T}\sum_{b=1}^{T} (\hat{\tau}_b - \bar{\tau})(\hat{\tau}_b - \bar{\tau})'.$$

Here $\bar{\tau}$ is the mean of $\hat{\tau}_b$ across bootstrap trials. We then compare $\hat{W}$ to critical values of the χ^2_t distribution, where t is the number of elements in $\hat{\tau}$. The results of this test are shown in Appendix Table A.VII.

Appendix G: Site Group Fixed Effects

This appendix describes methods for incorporating experimental site group fixed effects into our two-step control function estimation procedure. These methods allow us to leverage cross-site variation while reducing the dimension of heterogeneity across sites, eliminating an incidental parameters problem that would arise with a full set of site fixed effects. Our approach is similar in spirit to that of Bonhomme and Manresa (2015), who develop methods that account for grouped patterns of heterogeneity in linear panel data models. In the translation from panel data to our multi-site experimental setting, sites play the role of cross-sectional units and experimental subjects play the role of time periods.

G.1 Model

Experimental sites are indexed by $s \in \{1, .., S\}$, and $s(i)$ denotes the site for individual $i \in \{1, ..., N\}$. Each site belongs to one of G unobserved groups, with $g(s) \in \{1, ..., G\}$ the group for site s. The number of sites S may grow asymptotically with N, but the number of groups G is assumed to be fixed. Utilities for Head Start, other preschools and home care are given by

$$\begin{aligned} U_i(h, Z_i) &= \psi_h^{g(s(i))}(Z_i) + v_{ih}, \\ U_i(c) &= \psi_c^{g(s(i))} + v_{ic}, \\ U_i(n) &= 0, \end{aligned}$$

with

$$(v_{ih}, v_{ic}) \,|Z_i, s(i) \sim N\left(0, \begin{bmatrix} 1 & \rho^{g(s(i))} \\ \rho^{g(s(i))} & 1 \end{bmatrix}\right).$$

Here we have omitted other observed covariates for simplicity, though these can be easily incorporated. This model implies that preferences depend on the site $s(i)$ through the site group $g(s(i))$. This reduces the dimension of cross-site heterogeneity from S to G.

G.2 Estimation

If the site groupings were known, the group-specific parameters $\Psi = \left\{\psi_h^g(1), \psi_h^g(0), \psi_c^g, \rho^g\right\}_{g=1}^G$ could be straightforwardly estimated via a multinomial probit model saturated in group indicators. These groupings are unknown *a priori*, however, so the group assignments must be estimated from the data. Following Bonhomme and Manresa (2015), we use an estimation scheme that alternates between maximizing the likelihood function conditional on group assignments and reassigning groups to maximize the likelihood function conditional on the group-specific parameters.

Let $g_0(s)$ be the initial type assignment for site s. The estimated group-specific parameters at iteration $k \in \{0, 1, ...\}$ are given by

$$\hat{\Psi}_k = \arg\max_{\Psi} \sum_{i=1}^{N} \log \mathcal{L}\left(D_i | Z_i; \psi_h^{g_k(s(i))}(1), \psi_h^{g_k(s(i))}(0), \psi_c^{g_k(s(i))}, \rho^{g_k(s(i))}\right),$$

where $\mathcal{L}(d|z; \psi_h(1), \psi_h(0), \psi_c, \rho)$ is the multinomial probit likelihood function. Let $\left\{\hat{\psi}_h^{g_k}(1), \hat{\psi}_h^{g_k}(0), \hat{\psi}_c^{g_k}, \hat{\rho}^{g_k}\right\}$ denote the elements of $\hat{\Psi}_k$ corresponding to group g. The new group assignments for iteration $k+1$ are then

$$g_{k+1}(s) = \arg\max_{g \in \{1 \ldots G\}} \sum_{i:s(i)=s} \log \mathcal{L}\left(D_i | Z_i; \hat{\psi}_h^{g_k}(1), \hat{\psi}_h^{g_k}(0), \hat{\psi}_c^{g_k}, \hat{\rho}_k^{g_k}\right).$$

The algorithm proceeds until the change in the log likelihood from one iteration to the next falls below a tolerance threshold.

G.3 Implementation

Before implementing the estimation procedure, we group together very small sites until the remaining sites have no fewer than 10 observations. Where possible, sites with the smallest numbers of observations are first grouped together within Head Start program areas until the smallest site within an area has at least 10 observations (see Puma et al., 2010 for a description of HSIS program areas and experimental sites). For program areas with fewer than 10 total observations, we then iteratively group the smallest program areas into sites until the smallest site has no fewer than 10. This procedure results in 183 sites with average size 19.5.

The group fixed effects estimator described above is then applied to the sites. The objective function for the group fixed effects estimation procedure may not be globally concave. To aid in finding the global maximum, we sequentially increase the complexity of the model by estimating it for each G and using the final group assignments from the previous model to initialize the next model. Specifically, to estimate a model with G groups, we start with the final assignments from a model with $G-1$ groups and split the group with the lowest final log likelihood at the median log likelihood. This procedure performed well in Monte Carlo trials.

To avoid overfitting the model, we select the final number of groups based on the Bayesian Information Criterion (BIC). The BIC penalizes extra parameters in proportion to the log of the sample size. Let $g_G^*(s)$ denote the final group assignment for site s when the total number of groups is G. The BIC is given by

$$BIC(G) = -2\sum_{i=1}^{N} \log \mathcal{L}\left(D_i | Z_i; \hat{\psi}_h^{g_G^*(s(i))}(1), \hat{\psi}_h^{g_G^*(s(i))}(0), \hat{\psi}_c^{g_G^*(s(i))}, \hat{\rho}^{g_G^*(s(i))}\right) + (S + 4G) \log N.$$

Here the S in the second term captures parameters corresponding to group assignments for the S sites, while the $4G$ captures the estimated group-specific parameters. The final number of groups is chosen to minimize $BIC(G)$. As shown in Appendix Table A.VI, the BIC selects 7 groups when the model includes no other covariates and 6 groups when the model includes our full set of baseline covariates.

Our two-step models with site group fixed effects include indicators for site groups in all second-step regressions, fully interacted with preschool alternative. The site groups and group-specific parameters are reestimated in our bootstrap resampling procedure, with group assignments initialized at their full-sample values in each bootstrap trial.

Additional Appendix References

1. Abadie, A. (2002). "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variables Models." Journal of the American Statistical Association 97(457).
2. Ackerberg, D., and Devereux, P. (2009). "Improved JIVE Estimators for Overidentified Linear Models With and Without Heteroskedasticity." The Review of Economics and Statistics 91(2).
3. Angrist, J., Imbens, G., and Krueger, A. (1995). "Jackknife Instrumental Variables Estimation." NBER Working Paper no. 172.
4. Ashenfelter, O., Doran, K., and Schaller, B. (2010). "A Shred of Credible Evidence on the Long-run Elasticity of Labor Supply." Economica 77(308).
5. Blundell, R., Pistaferri, L., and Saporta-Eksten, I. (forthcoming). "Consumption inequality and family labor supply." American Economic Review.
6. Cascio, E., and Staiger, D. (2012). "Knowledge, Tests, and Fadeout in Educational Interventions." NBER Working Paper no. 18038.
7. Chao, J., Hausman, J., Newey, W., Swanson, N., and Woutersen, T. (2013). "Testing Overidentifying Restrictions With Many Instruments and Heteroskedasticity." Journal of Econometrics 178.
8. Currie, J. (2006). "The Take-up of Social Benefits." In Alan Auerbach, David Card, and John Quigley, eds., *Poverty, the Distribution of Income, and Public Policy.* New York, NY: The Russell Sage Foundation.
9. Hansen, L. (1982). "Large Sample Properties of Generalized Method of Moments Estimators." Econometrica 50(4).
10. Heckman, J., Moon, S., Pinto, R., Savelyev, P., and Yavitz, A. (2010b). "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program." Quantitative Economics 1(1).
11. Heckman, J., Stixrud, J., and Urzua, S. (2006). "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." Journal of Labor Economics 24(3).
12. Krueger, A. (2003). "Economic Considerations and Class Size." Economic Journal 113(485).
13. Lindqvist, E., and Vestman, R. (2011). "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment." American Economic Journal: Applied Economics 3(1).
14. Murnane, Willett and Levy (1995). "The Growing Importance of Cognitive Skills in Wage Determination." The Review of Economics and Statistics 77(2).
15. Sojourner, A. (2009). "Inference on Peer Effects With Missing Peer Data: Evidence from Project STAR." Working Paper.
16. Tallis, G. (1961). "The Moment Generating Function of the Truncated Multi-normal Distribution." Journal of the Royal Statistical Society 23(1).

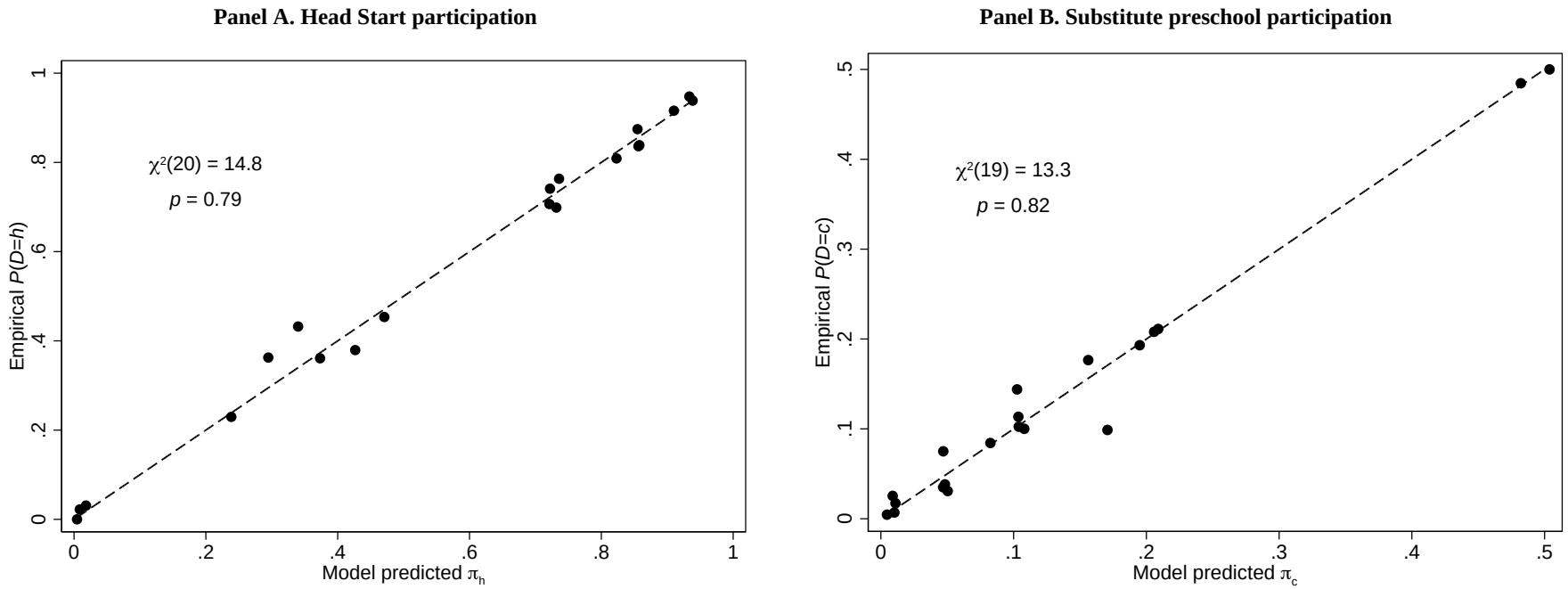

Figure A.I. Multinomial Probit Model Fit

Notes: This figure plots empirical probabilities of participating in Head Start and competing preschools against corresponding model predictions. Estimates come from the multinomial probit model in Table VI. Cells are defined by interactions of quintiles of the two predicted probabilities from the model. Cells with fewer than 50 observations are combined into a single cell. Panel A compares empirical probabilities of Head Start participation against cell means of the corresponding model-predicted probability, and panel B shows corresponding results for substitute preschools. Each panel shows the results of a test that the empirical and model-predicted probabilities are equal, treating the model predictions as fixed. The joint p-value for a test that the model fits in both panels equals 0.76.

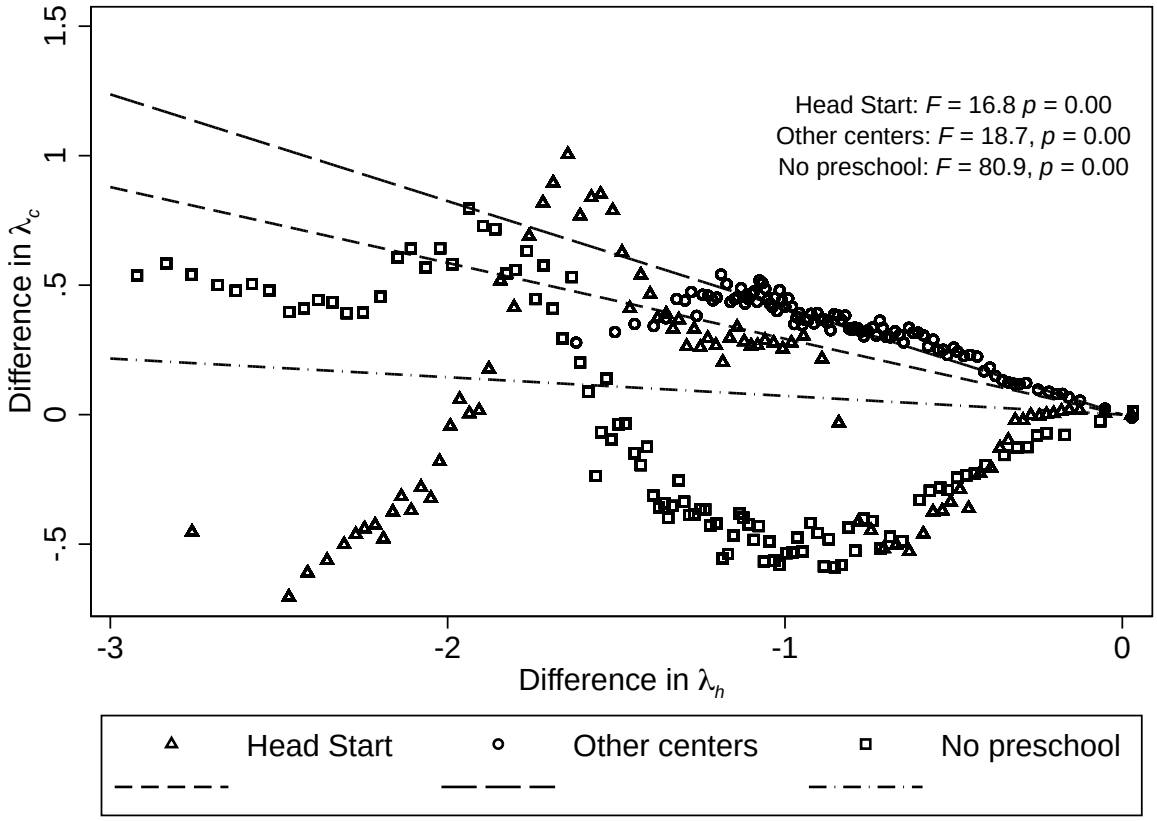

Figure A.II. Identification of the Selection Model

Notes: This figure plots differences in control functions that predict Head Start and other preschool tastes conditional on preschool choices and covariates. Estimates come from the multinomial probit model in Table VI. The horizontal axis shows the difference in predicted Head Start tastes with the Head Start offer switched on and off, and the vertical axis shows the difference in predicted other preschool tastes with the offer switched on and off. Identification of the selection model requires that these values do not all lie on a line through the origin for each preschool choice. Dashed lines show OLS fits through the origin, and points show means of control function differences by percentile of the difference in predicted Head Start tastes. Tests are based on regressions of the difference in λ_h on a constant and a third-order polynomial in the difference in predicted λ_c for each preschool choice. F-statistics and p-values come from bootstrapped Wald tests of the hypothesis that the constant, second- and third-order terms are zero. See Appendix F for details. To preserve scale, the figure omits points in the bottom decile of the predicted difference in tastes for Head Start.

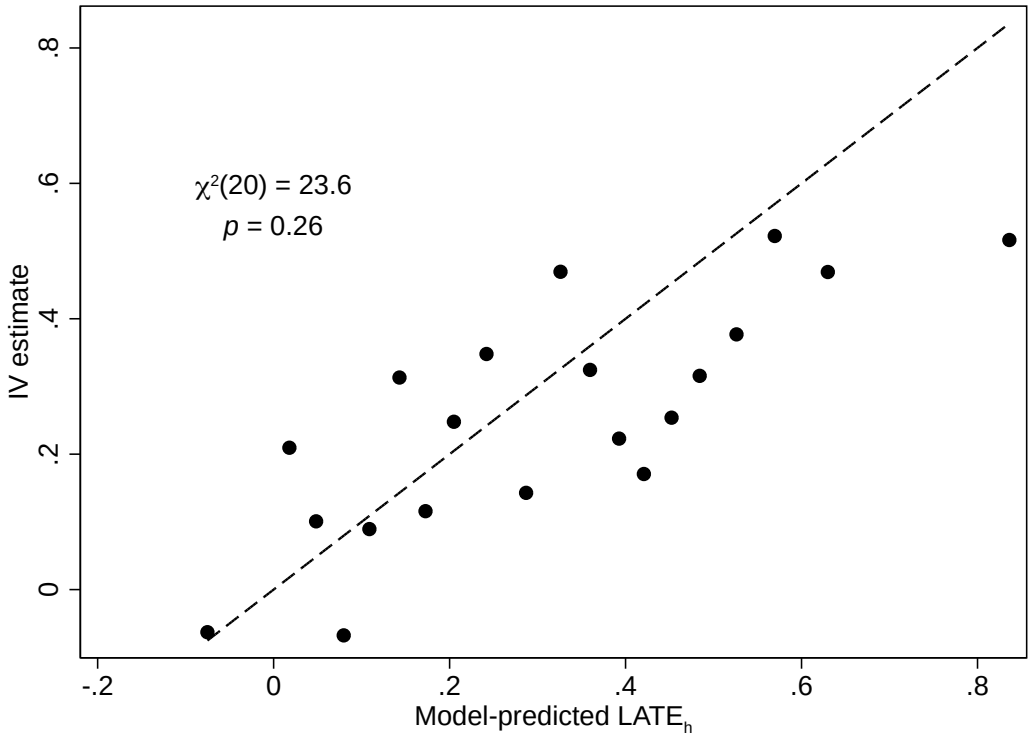

Figure A.III. Model-predicted $LATE_h$ vs. IV estimates

Notes: This figure plots model-predicted local average treatment effects against IV estimates. Estimates come from the two-step model in column (5) of Table VII. The sample is divided into vingtiles on the basis of the model-predicted LATE. Points show IV estimates by vingtile vs. average model-predicted LATE by vingtile. The dashed line is the 45 degree line. Test statistic and p-value come from a Wald test of the hypothesis that the 45 degree line fits all points up to sampling error.

Table A.I. Characteristics of Head Start Centers Attended by Always Takers

	Experimental center (1)	Attended center (2)
Transportation provided	0.421	0.458
Quality index	0.701	0.687
Fraction of staff with bachelor's degree	0.304	0.321
Fraction of staff with teaching license	0.084	0.099
Center director experience	19.08	18.24
Student/staff ratio	6.73	6.96
Full day service	0.750	0.715
More than three home visits per year	0.112	0.110
N	112	
p -value	0.318	

Notes: This table reports characteristics of Head Start centers for children assigned to the HSIS control group who attended Head Start. Column (1) shows characteristics of the centers of random assignment for these children, while column (2) shows characteristics of the centers they attended. The *p* -value is from a test of the hypothesis that all mean center charteristics are the same. The sample excludes children with missing values for either characteristics of the center of random assignment or the center attended.

Table A.II. Characteristics of Head Start and Substitute Preschool Centers

Panel A. Funding sources				**Panel B. Inputs and practices**			
Largest funding source	Head Start (1)	Other centers (2)	Other centers attended by $c \rightarrow h$ compliers (3)	Input	Head Start (4)	Other centers (5)	Other centers attended by $c \rightarrow h$ compliers (6)
Head Start	0.842	0.027	0.038	Transportation provided	0.629	0.383	0.324
Parent fees	0.004	0.153	0.191	Quality index	0.702	0.453	0.446
Child and adult care food program	0.011	0.026	0.019	Fraction of staff with bachelor's degree	0.345	0.527	0.491
State pre-K program	0.004	0.182	0.155	Fraction of staff with teaching license	0.113	0.260	0.247
Child care subsidies	0.013	0.097	0.107	Center director experience	18.2	12.2	12.6
Other funding or support	0.022	0.118	0.113	Student/staff ratio	6.80	8.24	8.54
No funding or support	0.000	0.003	0.001	Full day service	0.637	0.735	0.698
Missing	0.105	0.394	0.375	More than three home visits per year	0.192	0.073	0.072

Notes: This table reports characteristics of Head Start and other preschool centers obtained from surveys of center directors. Panel A displays information on the largest funding source for each center type, and panel B shows information on center inputs and practices. Columns (3) and (6) reports characteristics of other preschool centers attended by non-offered compliers who would be induced to attend Head Start by an experimental offer. Estimates in these columns are produced using the methods for characterizing compliers described in Appendix B.

Table A.III. Effects on Maternal Labor Supply

	Full-time	Full- or part-time
	(1)	(2)
Offer effect	0.020	-0.005
	(0.018)	(0.019)
Mean of dep. var.	0.334	0.501
N	3314	

Notes: This table reports coefficients from regressions of measures of maternal labor supply in Spring 2003 on the Head Start offer indicator. Column (1) displays effects on the probability of working full-time, while column (2) shows effects on the probability of working full- or part-time. Children with missing values for maternal employment are excluded. All models use inverse probability weights and control for baseline covariates. Standard errors are clustered at the Head Start center level.

Table A.IV. Estimates of Test Score and Earnings Impacts

Study	Intervention (1)	Test score effect (std. dev. units) (2)	Log earnings effect (3)	Log wage effect (4)	Ratio: wages or earnings /test scores (5)
Chetty et al. (2011)	Tennessee STAR (1 s.d. of class quality, kindergarten)[a]	0.024	0.003	-	0.131
	OLS with controls (kindergarten)[b]	1.0	0.18	-	0.18
Chetty et al. (2014b)	Teacher value-added (1 s.d. of teacher VA, grades 3-8)[c]	0.13	0.013	-	0.103
	OLS with controls (grades 3-8)[d]	1.0	0.12	-	0.12
Currie and Thomas (1999)	OLS with controls (age 7)[e]	1.0	-	Partial effects: 0.076 (math), 0.080 (reading)	0.076 (math), 0.080 (reading)
Currie and Thomas (1995), Garces et al. (2002)	Head Start (whites, mother fixed effects, age 4+)[f]	0.217	0.566	-	2.61
	Head Start (blacks, mother fixed effects, age 4+)[g]	0.009	0.073	-	8.11
Heckman et al. (2006)	OLS with controls (males, ages 14-22)[h]	1.0	-	0.121	0.121
	OLS with controls (females, ages 14-22)[i]	1.0	-	0.169	0.169
Heckman et al. (2010b)	Perry Preschool Project (males, age 4)[j]	0.787	0.189	-	0.240
	Perry Preschool Project (females, age 4)[k]	0.980	0.286	-	0.292
Lindqvist and Vestman (2011)	OLS with controls (males, w/controls, ages 18-19)[l]	1.0	0.136	0.104	0.104
Murnane et al. (1995)	OLS with controls (males, grade 12)[m]	1.0	-	0.077	0.077
	OLS with controls (females, grade 12)[n]	1.0	-	0.109	0.109

Notes: We convert all test score effects to standard deviation units (column (2)) and all earnings effects to percentages (column (3)).

[a]Table VIII: A 1 s.d. increase in class quality (peer scores) raises kindergarten test scores by 0.662 percentile points and age 27 earnings by $50.61.

[b]Table IV: Controlling for covariates, a 1 percentile point increase in kindergarten test scores raises average annual earnings from age 25 to age 27 by $93.79.

[c]Table III: A 1 s.d. increase in teacher value-added raises test scores by 0.13 standard deviations and boosts age 28 earnings by $285.55.

[d]Appendix Table III: Controlling for covariates, a 1 s.d. increase in test scores raises age 28 earnings by $2,585.

[e]Tables 3 and 4 report partial effects of scoring in the top vs. bottom quartile of reading and math scores at age 7 on log wages at age 33 for British children. We use Krueger's (2003) conversion of effects on quartiles to standard deviation units.

[f]Currie and Thomas (1995), Table 4: Head Start participation raises test scores by 5.88 percentile points at age 4+ for whites. Garces et al. (2002), Table 2: Head Start participation raises log earnings between age 23 and age 25 by 0.566 for whites.

[g]Currie and Thomas (1995), Table 4: Head Start participation raises test scores by 0.247 percentile points at age 4+ for whites. Garces et al. (2002), Table 2: Head Start participation raises log earnings between age 23 and age 25 by 0.073 for blacks.

[h]Table 1: Controlling for covariates, a one standard deviation increase in cognitive skills at age 14-22 increases log wages at age 30 by 0.121 for males. Controls include non-cognitive skills.

[i]Table 1: Controlling for covariates, a one standard deviation increase in cognitive skills at age 14-22 increases log wages at age 30 by 0.169 for females. Controls include non-cognitive skills.

[j]Appendix Figure G.1 (a): Treatment increased male IQ by 11.8 points at age 4. Appendix Table H.1: Treatment increased male age 27 earnings by $2,363 (control mean $12,495).

[k]Appendix Figure G.1 (b): Treatment increased female IQ by 14.7 points at age 4. Appendix Table H.2: Treatment increased female age 27 earnings by $2,568 (control mean $8,986).

[l]Table 1: Controlling for a small set of covariates, a one standard deviation increase in cognitive skills at age 18-19 increases log wages by 0.104 at age 32+ for Swedish men. Table 3: A one standard deviation increase in cognitive skills increases annual earnings by 43,392 SEK (sample mean 319,800 SEK).

[m]Table 3: Controlling for covariates, a 1-point increase in senior-year math scores increases age 24 log wages by 0.011 for males in the High School and Beyond Survey (the std. dev. of math scores is approximately 6.25 points).

[n]Table 4: Controlling for covariates, a 1-point increase in senior-year math scores increases age 24 log wages by 0.017 for females in the High School and Beyond Survey (the std. dev. of math scores is approximately 6.25 points).

Table A.V. Two Stage Least Squares Estimates with Site Interaction Instruments

		One endogenous variable	Two endogenous variables	
		Head Start	Head Start	Other centers
Instruments	Estimator	(1)	(2)	(3)
Offer	2SLS	0.247	-	-
(1 instrument)		(0.031)		
Offer x sites	2SLS	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	0.002	0.002	
	LIML	0.218	0.029	-0.581
		(0.027)	(0.139)	(0.432)
	Overid. *p*-value	0.002	0.076	
	JIVE	0.217	0.109	-0.329
		(0.026)	(0.110)	(0.332)
	Overid. *p*-value	0.001	0.003	

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 remaining rows instrument Head Start and other preschool attendance with interactions of the offer and indicators for experimental sites. Sites with fewer than 10 observations are grouped together within program areas as described in Appendix D. All models control for main effects of the interacting variables and baseline covariates. JIVE refers to the JIVE2 estimator defined in Angrist, Imbens and Krueger (1995), computed after first partialing out the exogenous covariates as described by Ackerberg and Devereux (2009). Overidentification tests for JIVE are based on Hansen's (1982) *J*-statistics for 2SLS and LIML. Overidentification tests for JIVE are based on the many instrument and heteroskedasticity-robust statistic derived by Chao et al. (2013). First stage *F*-statistics are Angrist/Pischke (2009) partial *F*'s. Standard errors are robust to heteroskedasticity.

Table A.VI. Model Selection Criteria for Site Group Fixed Effect Models

	Sites only		Covariates and sites	
	Log likelihood	BIC	Log likelihood	BIC
Groups	(1)	(2)	(3)	(4)
1	-2,761.7	7,323.1	-2,582.0	7,912.7
2	-2,535.0	6,657.1	-2,366.9	6,811.6
3	-2,435.6	6,490.9	-2,268.3	6,647.1
4	-2,386.9	6,426.4	-2,223.4	6,590.0
5	-2,348.5	6,382.2	-2,184.1	6,544.2
6	-2,309.0	6,336.0	-2,154.9	6,518.6
7	-2,292.2	6,335.0	-2,150.6	6,542.8
8	-2,279.1	6,341.7	-2,141.7	6,557.5

Notes: This table shows results for multinomial probit models with fixed effects for unobserved experimental site groups. Columns (1) and (3) show the maximized log likelihood for each number of site groups, and columns (2) and (4) show corresponding values of the Bayesian Information Criterion (BIC), equal to the number of model parameters times the log of the sample size minus twice the log likelihood. Columns (1) and (2) include no other covariates, while columns (3) and (4) include the covariates listed in the notes to Table VI. See Appendix G for details.

Table A.VII. Comparison of IV and Model-based Estimates of Mean Potential Outcomes

	Type probability		$E[Y(h)]$		$E[Y(c)]$		$E[Y(n)]$	
	IV	Two-step	IV	Two-step	IV	Two-step	IV	Two-step
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
n-compliers	0.454	0.454	-	0.303	-	-0.323	-0.078	-0.067
c-compliers	0.232	0.231	-	0.078	0.107	0.172	-	-0.525
All compliers	0.686	0.685	0.233	0.227	-	-0.156	-	-0.221
n-never takers	0.095	0.093	-	0.590	-	-0.392	-0.035	-0.017
c-never takers	0.083	0.082	-	0.248	0.316	0.309	-	-0.530
Always takers	0.136	0.140	-0.028	0.027	-	-0.140	-	-0.340
Full population	1	1	-		-	-0.136	-	-0.245
P-value: IV = Two-step	0.589		0.260		0.605		0.731	
P-value for all moments	0.792							

Notes: This table compares nonparametric estimates of mean potential outcomes for subpopulations to estimates implied by the two-step model in column (5) of Table VII.