# Evaluating Public Programs with Close Substitutes: The Case of Head Start

Patrick Kline
UC Berkeley/NBER

Christopher Walters*
UC Berkeley/NBER

May 2015

**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 that receive public funds. This both attenuates measured experimental impacts on test scores and reduces the program's net budgetary costs. A calibration exercise indicates that accounting for the public savings associated with reduced enrollment in other subsidized preschools substantially increases estimates of Head Start's rate of return, defined as the after-tax lifetime earnings generated by an extra dollar of public spending. Estimates of a semi-parametric selection model reveal substantial heterogeneity in Head Start's test score impacts with respect to counterfactual care alternatives as well as observed and unobserved child characteristics. Head Start is about as effective at raising test scores as competing preschools and its impacts are greatest on children from families unlikely to participate in the program. Expanding Head Start to new populations is therefore likely to boost the program's rate of return, provided that the proposed technology for increasing enrollment is not too costly.

Keywords: Program Evaluation, Head Start, Early-Childhood Education, Marginal Treatment Effects

*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 Meetings, Columbia University, and UC San Diego for helpful comments and Raffaele Saggio for providing 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.

# 1 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. In particular, neglecting program substitution can lead to overstatement of a program's budgetary costs if competing programs are also publicly financed. This paper assesses the cost-effectiveness of Head Start – a prominent 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, Head Start 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 (US DHHS 2010, 2012a). 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, recent research on early childhood interventions finds long run gains in adult outcomes despite short run fadeout of test scores impacts (Heckman et al., 2010a, 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 whether they received 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 statistically imprecise impact estimates 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. Surveys of center administrators indicate that these compliers would have attended competing preschool programs that draw heavily on public funding, which mitigates the net costs to government of enrolling them in 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 a social planner seeks to maximize the after-tax lifetime earnings of a cohort of children subject to a budget constraint. We show that, for purposes of determining optimal program scale, the policy-relevant causal parameter governing program benefits is an average effect of Head Start participation on test scores 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 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. 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 or marketing efforts targeting children who are unlikely to attend. 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 causal 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 semi-parametric selection model that parameterizes treatment effect heterogeneity with respect to counterfactual care alternatives as well as observed and unobserved child characteristics. We provide a constructive proof of identification of the model parameters and show that our estimates accurately reproduce patterns of treatment effect heterogeneity found in the experiment. The estimated parameters 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 small effects for 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. These results suggest that expanding the program to new populations can boost its effectiveness.

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 substantially 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 seats in competing preschool programs vacated by Head Start compliers. Our estimates indicate that the effect on test scores (and therefore earnings) of moving eligible 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 calibration analysis.

The rest of the paper is structured as follows. Section 2 provides background on Head Start. Section 3 describes the HSIS data and basic experimental impacts. Section 4 presents evidence on substitution patterns. Section 5 introduces a theoretical framework for assessing public programs with close substitutes. Section 6 provides a cost-benefit analysis of Head Start. Section 7 develops our econometric selection model and discusses identification and estimation. Section 8 reports estimates of the model. Section 9 simulates the effects of structural program reforms. Section 10 concludes.

## 2 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). The program is also oversubscribed. In 2002, 85 percent of Head Start participants attended programs with more applicants than available seats (US DHHS, 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, 2012b). 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, 2009; 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 (US DHHS 2010, 2012a). 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 com-

[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 (forthcoming) 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.

parisons. 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. Participation rates in other public preschool programs have risen dramatically over time, so alternative preschool options were likely more accessible for HSIS applicants than for earlier cohorts of Head Start participants (Cascio and Schanzenbach, 2013). Indeed, many children in the HSIS control group attended other public or private preschools, and some children in the treatment group declined the Head Start offer in favor of other preschools. One aim of our analysis is to clarify the impact of this program substitution on the results of the HSIS experiment.

# 3 Data and Experimental Impacts

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

### 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 1 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 1 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 1 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 among applicants offered a slot in Head Start were similar to those denied a slot.[2]

Columns (3) through (5) of Table 1 report summary statistics for children attending Head Start, other preschool centers, and no preschool, respectively. Children in non-Head Start preschools

[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 advanced, this is an error-free measure of the probability of an offer for most applicants (DHHS 2010).

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

### Experimental impacts on test scores

Table 2 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 1.[3] The intent-to-treat estimates mirror those previously reported in the literature (e.g., US DHHS, 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 0.04 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 (0.039 standard deviations).

Interpretation of these intent-to-treat impacts is clouded by noncompliance with random assignment. Columns (2), (5) and (8) of Table 2 report first-stage effects of assignment to Head Start on the probability of participating in Head Start.[4] Columns (3), (6) and (9) report two-stage least squares (2SLS) estimates, which scale the intent-to-treat by the first stage. 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 (Imbens and Angrist 1994). Assignment to Head Start increases the probability of participation by two-thirds in the first year after random assignment. The corresponding 2SLS 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

[3]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, special education, test language, home language, dummies for quartiles of family income and missing income, an indicator for whether the Head Start center provides transportation, the Head Start quality index, and a third-order polynomial in baseline test scores.

[4]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 A1 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.

reapply for Head Start, resulting in substantially 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; the intent-to-treat falls roughly in proportion, generating a second-year 2SLS estimate of 0.249 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 0.114 standard deviations, with a standard error of 0.097. The corresponding first grade estimate for four year olds is 0.082 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 relatively 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).[5]

# 4 Substitution patterns in the HSIS

We now turn to documenting program substitution patterns in the HSIS and their potential influence on the HSIS results. It is helpful to develop some notation to describe the role of alternative care environments in the HSIS. 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 $D_i \in \{h, c, n\}$ denote household $i$'s treatment choice. Treatment choices may be affected by the experimental Head Start offer. Let $Z_i \in \{0, 1\}$ indicate whether household $i$ has a Head Start offer, and $D_i(z)$ denote 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 their behavior as a result

[5]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 (DHHS, 2010; Walters, forthcoming). The HSIS non-cognitive outcomes differ significantly from those analyzed in previous studies, however, and it is unclear whether they capture the same skills.

of the Head Start offer does so to attend Head Start.

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

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

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

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: $D_i(1) = D_i(0) = h$.

$n$- 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.

Figure 1 presents a graphical representation of compliance patterns in the HSIS. 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; as a result, the set of children enrolled in Head Start is a mixture of always takers and the two groups of offered compliers.

Table 3 presents corresponding 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.

How do the compliance patterns displayed in Table 3 affect the interpretation of the HSIS test score impacts? Formally, 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 $Y_i = Y_i(D_i)$. Assume that Head Start offers affect test scores only through program participation choices, an exclusion restriction that seems plausible in this context. Under assumption (1), instrumental variables (IV) estimation 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 the relevant set of counterfactual alternatives. Specifically, it is straightforward to show that under (1) and excludability of Head

Start offers:

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

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 2SLS strategy employed in Table 2 yields the average effect of Head Start for compliers relative to their own counterfactual preschool 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

$$\begin{aligned} LATE_{dh} &\equiv E\left[Y_i(h)-Y_i(d)|D_i(1)=h, D_i(0)=d\right],\ d\in\{c,n\}, \\ S_c &\equiv \frac{P\left(D_i\left(1\right)=h, D_i\left(0\right)=c\right)}{P\left(D_i\left(1\right)=h, D_i\left(0\right)\neq h\right)}. \end{aligned} \tag{4}$$

The weight $S_c$ gives the fraction of compliers drawn from other preschools, which can be estimated using the empirical analogue of the following equation:

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

Column (7) of Table 3 shows that 34 percent of compliers would have otherwise attended competing preschools in the first year of the HSIS experiment. Our 2SLS estimates combine effects for these compliers with effects for compliers who would not otherwise attend preschool. We next consider the implications of this fact for assessments of Head Start's cost-effectiveness.

# 5 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 of Head Start 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. The analysis ignores redistributional motives and any effects of human capital investment on criminal activity (Lochner and Moretti, 2004; Heckman et al., 2010a), health (Deming, 2009; Carneiro and Ginja, 2014), or grade repetition (Currie, 2001). Adding such features would tend to raise the implied return to Head Start. We also abstract from parental labor supply decisions, both because prior analyses of the Head Start Impact Study find no impact on parental labor supply decisions (US DHHS 2010, 2012a)[6] and because some research suggests that Head Start actually crowds-in parental time investment (Gelber and Isen, 2013). Again, incorporating these adjustments would likely raise the program's rate of return.

To mirror the nature of the public debate over early education programs, we work with a money metric social criterion where a planner seeks to maximize the after-tax lifetime revenue of a cohort of children. We show in Appendix C however that our proposed rate of return metric can be justified as an approximation to a more complicated intergenerational problem where a planner seeks to maximize steady state household utilities. Importantly, both models have the novel feature of accommodating unrestricted treatment effect heterogeneity, which allows us to be precise about how policy relevant "sufficient statistics" (Saez, 2002; Chetty, 2009; Hendren, 2014) map into particular causal estimands of interest. Specifically, we show that a variant of the Local Average Treatment Effect (LATE) concept of Imbens and Angrist (1994) is policy relevant when considering program expansions, while a variant of the Marginal Treatment Effect (MTE) of Heckman and Vytlacil (1999) is relevant when considering reforms to Head Start program features.

### Setup

Our model begins by motivating the behavioral restrictions in (1) via utility maximization. Consider a continuum 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.

Each household $i$ has utility over its enrollment options given by the function $U_i(d, z)$. The argument $d \in \{h, c, n\}$ indexes enrollment options, 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:

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

The valuations $\{U_i(h, 1), U_i(h, 0), U_i(c), U_i(n)\}$ are distributed according to a differentiable joint distribution function $F_U(., ., ., .)$.

Households make enrollment decisions by maximizing utility conditional on offer status. Hence, household $i$'s conditional enrollment function is determined by:

$$D_i(z) = \underset{d \in \{h,c,n\}}{\arg\max}\, U_i(d, z). \tag{5}$$

It is straightforward to show that this model generates the restrictions on substitution patterns displayed in Figure 1. Specifically, the assumption that offers do not affect utilities for $c$ and

[6] We replicate this analysis for our sample in Table A2. This Table 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.

$n$ implies the monotonicity restriction (1). As before, realized enrollment choices are given by $D_i = D_i(Z_i)$. Since offers are assigned at random, market shares can be written $P(D_i = d) = \delta P(D_i(1) = d) + (1-\delta) P(D_i(0) = d)$.

### A paternalistic planning criterion

Debate over the effectiveness of educational programs often centers on their test score impacts. This arguably reflects both a paternalistic emphasis on academic achievement and a belief (corroborated by recent research) that short-run test score impacts are linked to long-run effects on earnings and other adult outcomes (Heckman et al., 2010a, 2013; Chetty et al., 2011, 2014b). To characterize this view, we assume that society values test score outcomes according to the money metric criterion function:

$$W = (1-\tau) pE[Y_i], \tag{6}$$

where $p$ gives the market price of human capital and $\tau$ is the tax rate faced by the children of eligible households. That is, we assume society values schooling outcomes according to their impact on the after-tax income children realize as adults.

By writing (6) in terms of average test scores, we have neglected distributional impacts and the fact that Head Start provides a free in-kind transfer to poor households. In Appendix C, we study a utilitarian social objective in an overlapping generations model and show that it yields similar conclusions regarding the conditions necessary for optimality. Specifically, we show that the expressions derived here provide a lower bound to the utility-based rate of return experienced by households. In addition to being simpler, (6) has the advantage of reflecting the paternalistic nature of the debate over early education programs, which tends to focus on the cost of boosting the achievement of poor children while ignoring any impacts on adults.

We assume the planner faces the following budget constraint:

$$\phi_h P(D_i = h) + \phi_c P(D_i = c) = \tau pE[Y_i] + R. \tag{7}$$

The scalar $\phi_h$ gives the administrative cost of providing Head Start services to an additional child. Likewise, $\phi_c$ gives the administrative cost to governments of providing competing preschool services (which often receive subsidies) to another student. The term $\tau pE[Y_i]$ captures the revenue generated by taxes on the earnings of Head Start-eligible children, and $R$ reflects revenue received from other sources. 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.

### Optimal program scale

We assume the planner can adjust the scale of Head Start by changing the rationing probability $\delta$. An increase in $\delta$ induces additional compliers to attend Head Start, and these compliers are drawn both from competing programs and home care. As shown in Appendix B, the effect of a change in the offer rate $\delta$ on average test scores is given by:

$$\frac{\partial E\left[Y_i\right]}{\partial \delta} = LATE_h P\left(D_i\left(1\right)=h, D_i\left(0\right)\neq h\right). \tag{8}$$

This result follows from the assumption that Head Start offers are distributed at random and that $\delta$ 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 $\delta$.

As shown in Appendix B, the first order condition for maximization of (6) subject to (7) can be written:

$$\underbrace{\left(1-\tau\right)pLATE_h}_{\text{Income Gain}} = \underbrace{\phi_h}_{\text{Provision Cost}} - \underbrace{\phi_c S_c}_{\text{Public Savings}} - \underbrace{\tau pLATE_h}_{\text{Added Revenue}}, \tag{9}$$

The left-hand side of (9) gives the benefit of enrolling another child in Head Start. An average program complier experiences a test score gain of $LATE_h$, and a corresponding increase in after-tax earnings of $\left(1-\tau\right)pLATE_h$. The right hand side of (9) gives the net cost to government of adding another Head Start slot. This is the administrative cost $\phi_h$ minus the probability $S_c$ that the complying household comes from another preschool times the expected government savings $\phi_c$ associated with reduced enrollment in competing preschools. Subtracted from this cost is any extra revenue the government gets from raising the productivity of the children of complying households.

Adapting the arguments of Mayshar (1990) and Hendren (2014) to our paternalistic setting, we can take the ratio of the impact on after-tax income to the impact on the government budget constraint to obtain the marginal value of public funds (MVPF) associated with a change in the offer probability $\delta$:

$$MVPF_\delta = \frac{\left(1-\tau\right)pLATE_h}{\phi_h - \phi_c S_c - \tau pLATE_h}. \tag{10}$$

The MVPF 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. Appendix C pursues this intuition more formally.

An important lesson of the above analysis is that a cost-benefit analysis does not require identification of treatment effect heterogeneity with respect to the counterfactual care state. Specifically, it is not necessary to separately identify the subLATEs for the purpose of setting optimal program scale. 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 considering 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.

### Structural reforms

An important assumption of the previous analysis is that the program can be scaled up or down solely by changing lottery probabilities that do not alter the mix of program compliers. Suppose now that the planner can change enrollment by 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 might draw in households from a different mix of counterfactual care environments. By shifting the composition of program participants, changes in $f$ might (or might not) boost the program's rate of return.

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

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

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. As shown in Appendix B, the assumption that $f$ has no effect on potential outcomes implies:

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

where

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

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 (see Appendix B). 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 (8), 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 planner must balance the test score effects of improvements to the program feature against the costs. We suppose that changing program features changes the *average* cost of Head Start services, so that the government's budget constraint is now:

$$\phi_h(f)\,P(D_i=h)+\phi_c P(D_i=c)=\tau pE[Y_i]+R, \tag{11}$$

where $\phi_h'(f)\geq 0$.

An (interior) optimum to the planning problem ensues when:

$$\begin{aligned}\underbrace{(1-\tau)\,pMTE_h}_{\text{Income Gain}} = \underbrace{\phi_h}_{\text{Marginal Provision Cost}} + \underbrace{\phi_h'(f)\left(\partial\ln P(D_i=h)/\partial f\right)^{-1}}_{\text{Inframarginal Provision Cost}} \\ - \underbrace{\phi_c \overrightarrow{S}_c}_{\text{Public Savings}} - \underbrace{\tau pMTE_h}_{\text{Added Revenue}}.\end{aligned} \tag{12}$$

The left hand side of condition (12) gives the dollar value of the expected gain in average test scores associated with enrolling a marginal household into Head Start. The right hand side of condition (12) gives the program costs. The first term is 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:

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

As in our analysis of optimal program scale, equation (12) 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.

## 6 Cost-Benefit Analysis

We next use the HSIS data to conduct a formal cost-benefit analysis of changes to Head Start's program scale. This exercise requires estimates of each term in equation (9). We estimate $LATE_h$ and $S_c$ from the HSIS, and calibrate the remaining parameters using estimates in the literature. Calibrated parameters are listed in panel A of Table 6. 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. As we shall see, this lower bound estimate yields a surprisingly favorable assessment of the program.

### Representativeness of the HSIS data

The HSIS data are a nationally-representative random sample of Head Start applicants, and HSIS offers are distributed randomly (US DHHS, 2010). The HSIS is therefore ideal for estimating values of $LATE_h$ and $S_c$ *in the population of Head Start applicants.* Fortunately, the current Head Start application rate is extremely high, which limits the scope for selection into the applicant pool to 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 (US DHHS, 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 (9) gives the dollar value of a one standard deviation increase in test scores. Since earnings are unavailable for the HSIS sample, we project earnings impacts using literature estimates that relate test score and earnings impacts for educational interventions. Our approach is similar to that of Krueger (1999), who used estimates from Murnane et al. (1995) to predict earnings gains for the Tennessee STAR class size experiment prior to the availability of earnings data for this sample. A projection using this approach will produce reliable estimates of earnings impacts if the relationship between test score gains and earnings gains for Head Start is the same as this relationship for other programs, a strong assumption that we cannot test. We note, however, that a growing body of evidence shows a consistent link between short-run test score effects and earnings impacts.

Appendix Table A3 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.[7] The STAR results also 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 $LATE_h$.

[7] If test score distributions differ across populations or over time, effects in standard deviation units may have different meanings. For example, Cascio and Staiger (2012) show that norming in standard deviation units partially explains the phenomenon of test score fadeout in educational interventions. Sojourner (2009) shows that the standard deviation of nationally-normed percentile 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, respectively, for a mean of 81 percent. This suggests we should rescale the STAR estimate of 13.1 percent to roughly 12.2 percent in our sample; our baseline calibrations use a more conservative estimate of 10 percent.

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.[8] In an analysis of the Perry Preschool Project, Heckman et al. (2010) estimate larger ratios of earnings per standard deviation of test scores (24 to 29 percent). Mother 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 A3.[9]

Calculating percentage changes in earning 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, which suggests that 60 percent of the gap between Head Start families and the US average will be closed in their children's generation.[10] This implies that the average child in Head Start is expected to earn 78 percent of the US average $(1-(1-0.46)\cdot 0.4)$, which implies a present value of earnings of $\bar{e} = \$343,492$ at age 3.4.

Thus, we calculate that the marginal benefit of additional Head Start enrollment to be $0.1 \cdot \$343,492 \cdot LATE_h$. Using the pooled first-year estimate of $LATE_h$ reported in Section 3, 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.

[8] See also Havnes and Mogstad (2011) who find long run effects of subsidized childcare in Norway.

[9] The only estimate below 10 percent in Table A3 is from Murnane et al. (1995), who 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 also produces a larger estimate for females (10.9 percent).

[10] Chetty et al. (2014) find that the intergenerational income elasticity is not constant across the parent income distribution. Online 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 mean parent income but lower for parent incomes 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 somewhat below the figure of 0.4 used in our calculations. This in turn implies that our rate of return calculations are conservative in the sense that they underestimate the earnings impact of Head Start's test score gains.

### Program costs

Equation (9) 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 (DHHS, 2013). As reported in Column (7) of Table 3, the estimated share of compliers drawn from other preschools is 0.34. To investigate the cost to government of competing preschools, Table 4 reports information on funding sources for Head Start and other 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 D). 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.

Table 5 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 (1) and (2) show that Head Start centers are more likely to provide transportation 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 (3) shows that alternative preschools attended by Head Start compliers are very similar to the larger set of alternative preschools in the HSIS sample.

Taken together, the statistics in Tables 4 and 5 show that alternative preschool programs often involve public subsidies that are likely to generate social costs similar to Head Start. We conduct cost-benefit analyses under three assumptions: $\phi_c$ is either zero, 50 percent, or 75 percent of $\phi_h$. Table 4 suggests that roughly 75 percent of competing programs are financed primarily by public subsidies, so our preferred calculation uses $\phi_c = 0.75\phi_h$.

### Marginal value of public funds

Panel B of Table 6 reports estimates of the marginal value of public funds for Head Start expansion, $MVPF_\delta$. 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.[11] This approach can be shown to yield a higher-order refinement to $p$-values based upon the delta method normal approximation (Hall, 1992).

The results show that accounting for the public savings associated with enrollment in competing preschools has a large effect on the estimated social value of Head Start. Setting $\phi_c = 0$ yields a $MVPF$ of 1.10. Setting $\phi_c$ equal to $0.5\phi_h$ and $0.75\phi_h$ raises the $MVPF$ 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 $MVPF$ 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 $MVPF$'s for comparable expenditure programs summarized in Hendren (2014, Table 1). For example, Hendren computes values of $MVPF$ between 0.53 and 0.66 for food stamps, and 0.79 for public housing. Our preferred estimate of $MVPF$ for Head Start is 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 6 also reports "breakeven" values of $p/\bar{e}$ that set $MVPF$ equal to one for each value of $\phi_c$. When $\phi_c = 0$ the breakeven earnings effect is 9 percent, 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 $MVPF$ below one. Increasing $\phi_c$ to $0.5\phi_h$ or $0.75\phi_h$ reduces the breakeven earnings effect to 8 percent or 7 percent, respectively. The latter is less than all of the estimates in Table A3 and well below the Chetty et al. (2011) benchmark estimate of 13 percent. 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.

# 7 Econometric Model

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 ease of access, which could in turn change the mix of program compliers. To evaluate such reforms, it is necessary to predict how selection into the program is likely to change and how this impacts the program's rate of return.

To accomplish this goal, we now develop an econometric model geared toward characterizing the substitution patterns present in the HSIS experiment and their link to test score outcomes. The model allows us to conduct three analyses that are not possible with the basic experimental impacts. First, we separately estimate the "subLATEs" in equation (3) giving average impacts of Head Start on subgroups of compliers drawn from other preschools and home care. Substantial

[11] This test is computed by a non-parametric block bootstrap of the studentized $t$-statistic that resamples Head Start sites.

subLATE heterogeneity could explain why prior observational studies that compare home care to Head Start find impacts larger than those present in the HSIS. The subLATEs also indicate how the effects of Head Start might change if the program were altered to target children who would not otherwise attend preschool. Second, we compute social returns to reforms that expand Head Start by increasing its attractiveness (e.g. by providing transportation services) rather than simply changing offer probabilities. Finally, we conduct a cost/benefit analysis that extends the model of Section 5 to accommodate rationing of competing public preschool programs. Rationing of competing programs changes the expression for the marginal value of public funds and requires knowledge of additional parameters beyond $LATE_h$ and $S_c$.

### Identification problem

Our modeling approach leverages restrictions on selection patterns to impute mean potential outcomes that are not directly identified by the HSIS experiment. To understand the identification challenges posed by the presence of multiple treatment options, consider Table 7, which lists estimates of the mean potential outcomes nonparametrically identified by the HSIS. These estimates were computed using a generalization of results in Abadie (2002) to the case with multiple fallback treatments, which we detail in Appendix D. What follows is an informal discussion of the problem at hand.

Column (1) of Table 7 displays the population fractions of the five compliance groups discussed in Section 4, while Columns (3) through (5) report means of the three potential outcomes $Y_i(h)$, $Y_i(c)$ and $Y_i(n)$. Blank cells indicate means that are not identified. We can identify the mean of $Y_i(n)$ for $n$-never takers because observations with $Z_i = 1$ and $D_i = n$ are a random sample from this population. Likewise, we can identify the mean of $Y_i(c)$ for $c$-never takers from the population with $Z_i = 1$ and $D_i = c$, and we can identify the mean of $Y_i(h)$ for always takers from the population with $Z_i = 0$ and $D_i = h$. Since these three groups never change treatment status, their mean outcomes in other treatment states are not identified.

Mean potential outcomes are identified for the two complier subgroups in their preferred counterfactual states. Observed outcomes when $Z_i = 0$ and $D_i = n$ give a mixture of $Y_i(n)$ for $n$-never takers and $n$-compliers. Since the $n$-never taker mean and the relative frequencies of these two groups are identified, the mean of $Y_i(n)$ for $n$-compliers can be extracted from the observed mixture. A similar argument shows that the mean of $Y_i(c)$ for $c$-compliers can be recovered by removing the $c$-never taker mean from the group with $Z_i = 0$ and $D_i = c$, which mixes $c$-never takers and $c$-compliers.

Computing subLATEs requires computing the mean of $Y_i(h)$ for each complier subgroup. However, these means are not identified without further assumptions. To see this, note that the group with $D_i = h$ and $Z_i = 1$ is a mixture of always takers, $c$-compliers, and $n$-compliers. The mean of $Y_i(h)$ for the pooled set of compliers can be recovered by removing the always taker mean, but this leaves a mixture of the two complier subgroups. Identification of effects relative to specific alternatives requires assumptions that allow means of $Y_i(h)$ for complier subgroups to be extracted from the pooled complier mixture.

One approach to extracting the complier means is to conduct 2SLS estimation treating Head Start enrollment and enrollment in other preschools as separate endogenous variables. A common strategy for generating additional instruments in settings with multiple endogenous variables is to interact a single instrument with observed covariates or site indicators (e.g., Kling, Liebman and Katz, 2007; Abdulkadiroglu et al., 2014). Intuitively, such approaches presume that the instruments differ in their reduced form effects on outcomes only because of differences in their effects on the endogenous variables. We formalize this intuition in Appendix E, which shows that interacted 2SLS approaches will only identify subLATEs in our setting if covariate groups differ in their compliance shares $S_c$ but not their respective group-specific subLATEs. Appendix Table A4 reports estimates from this interacted 2SLS approach using the same key covariates used in the selection model to follow. The overidentifying restrictions in these models are rejected, indicating the presence of effect heterogeneity not explained by counterfactual care choices.

A second approach to the identification problem, based on the principal stratification framework of Frangakis and Rubin (2002), uses a parametric model to deconvolve the pooled complier mixture into distributions for subgroups. Feller et al. (2014) apply this method to estimate effects of Head Start relative to other preschools and home care in the HSIS. They assume normal distributions for potential outcomes within each of the five compliance groups. Under this assumption any non-normality in the pooled complier distribution must be due to differences between distributions for $n$- and $c$-compliers, which allows the components of the mixture to be recovered. Their results show large effects for $n$-compliers and negligible effects for $c$-compliers.

For our purposes a problem with both principal stratification and the interacted IV approach is that these methods condition on realized selection patterns and therefore cannot be used to predict the effects of reforms that change the mix of compliers. The empirical results in Table 7 show important differences between subgroups which imply non-random selection into care alternatives. For example, $c$-never takers have much higher scores than $c$-compliers in other preschools, and always takers perform worse than compliers in Head Start. Since current compliers differ systematically from other groups, we expect changes in the composition of compliers to generate changes in treatment effects. We next outline an econometric selection model that allows us to predict such changes.

## Selection model

Our selection model parametrizes the preferences and potential outcomes introduced in the model of Section 5. Normalizing the value of preschool non-participation to zero, we assume households have utilities over program alternatives given by:

$$\begin{aligned} U_i(h, Z_i) &= \psi_h^0 + X_i'\psi_h^x + \psi_h^z \cdot Z_i + Z_i \cdot X_i^{1\prime}\psi_h^{zx} + v_{ih}, \\ U_i(c) &= \psi_c^0 + X_i'\psi_c^x + v_{ic}, \\ U_i(n) &= 0, \end{aligned} \tag{13}$$

where $X_i = \left[X_i^1, X_i^2\right]$ denotes a vector of baseline household and experimental site characteristics and $X_i^1$ denotes a subset of these characteristics that we expect to shift the fraction of Head Start compliers who come from competing programs. In practice, $X_i^1$ consists of indicators for whether a child's center of random assignment offers transportation, above-median center quality, mother's education, age four, and an indicator for family income above the poverty line.

We use a multinomial probit specification for the stochastic components of utility:

$$\left(v_{ih}, v_{ic}\right)|X_i, Z_i \sim N\left(0, \left[\begin{array}{cc} 1 & \rho\left(X_i^1\right) \\ \rho\left(X_i^1\right) & 1 \end{array}\right]\right),$$

which allows for violations of the Independence from Irrelevant Alternatives (IIA) condition that underlies classic multinomial logit selection models such as that of Dubin and McFadden (1984). The correlation across alternatives is parameterized as follows:

$$\tanh^{-1}\left(\rho\left(X_i^1\right)\right) = \frac{1}{2}\ln\left(\frac{1+\rho\left(X_i^1\right)}{1-\rho\left(X_i^1\right)}\right) = \alpha^0 + X_i^{1\prime}\alpha^x.$$

By allowing both the error correlation and the effect of the Head Start offer on utility to vary with $X_i^1$, the model can accommodate rich heterogeneity in program substitution patterns.

To model endogeneity in participation decisions, we allow for a linear dependence of mean potential outcomes on the selection errors $(v_{ih}, v_{ic})$. Specifically, for each program alternative $d \in \{h, c, n\}$, we assume:

$$E\left[Y_i\left(d\right)|X_i, Z_i, v_{ih}, v_{ic}\right] = \theta_d^0 + X_i'\theta_d^x + \gamma_d^h v_{ih} + \gamma_d^c v_{ic}. \tag{14}$$

Assumption (14) can be thought of as a multivariate extension of the canonical Heckman (1979) sample selection model. While this approach is traditionally motivated by a joint normality assumption on the outcome and selection errors, (14) actually accommodates a wide variety of data generating processes exhibiting conditional heteroscedasticity and non-normality.[12]

The $\left\{\gamma_d^h, \gamma_d^c\right\}$ terms capture "essential" heterogeneity: treatment effect heterogeneity that is related to selection into treatment (Heckman, Urzua, and Vytlacil, 2006). Note that this specification can accommodate a variety of selection schemes. For example, if $\gamma_h^h = -\gamma_n^h$ then households engage in Roy (1951)-style selection into Head Start based upon test score *gains*. By contrast, if $\gamma_d^h = \gamma^h$ then selection into Head Start is governed by potential outcome *levels*.

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

[12]For example, the conditional distribution of potential outcomes could be a location-scale mixture of $K$ normal components with density $f_d\left(y\right) = \sum_{k=1}^{K}\frac{1}{\sigma_{dk}(X_i)}\tilde{\phi}\left(\frac{y-\theta_{dk}^0-X_i'\theta_{dk}^x-\gamma_{dk}^h v_{ih}-\gamma_{dk}^c v_{ic}}{\sigma_{dk}(X_i)}\right)\tilde{\pi}_{dk}\left(X_i\right)$, where $\tilde{\phi}$ is the standard normal density, $\sigma_{dk}\left(X_i\right)$ is a conditional variance function, and $\left\{\tilde{\pi}_{dk}\left(X_i\right)\right\}_{k=1}^{K}$ is a set of mixing weights which may depend on the covariates $X_i$ and the alternative $d$. As $K \to \infty$ this distribution can approximate any marginal distribution of potential outcomes (see Theorem 33.2 of DasGupta, 2008) . It is straightforward to verify that this model obeys (14) with $\gamma_d^h = \sum_{k=1}^{K}\gamma_{dk}^h E\left[\tilde{\pi}_k\left(X_i\right)\right]$ and $\gamma_d^c = \sum_{k=1}^{K}\gamma_{dk}^c E\left[\tilde{\pi}_k\left(X_i\right)\right]$.

written:

$$E\left[Y_i|X_i, Z_i, D_i = d\right] = \theta_d^0 + X_i'\theta_d^x + \gamma_d^h \lambda_{dh}\left(X_i, Z_i\right) + \gamma_d^c \lambda_{dc}\left(X_i, Z_i\right), \tag{15}$$

where $\lambda_{dh}\left(X_i, Z_i\right) \equiv E\left[v_{ih}|X_i, Z_i, D_i = d\right]$ and $\lambda_{dc}\left(X_i, Z_i\right) \equiv E\left[v_{ic}|X_i, Z_i, D_i = d\right]$ are multivariate generalizations of the standard inverse Mills correction term used in the Heckman (1979) selection framework. Appendix F provides analytical expressions for the selection correction terms using formulas for truncated bivariate normal integrals derived in Tallis (1961).[13] Intuitively, the selection correction terms are mean imputations for the unobservables driving endogeneity in the model. In a linear model, controlling for mean imputations of the unobservables is enough to remove their influence; hence, their moniker as "control functions."

### Identification of the selection model

It is straightforward to demonstrate identification of the selection coefficients $\{\gamma_d^h, \gamma_d^c\}$ from comparisons across covariate values of mean potential outcomes for the five compliance groups listed in Table 7. Consider the selection coefficients $\gamma_n^h$ and $\gamma_n^c$. Conditional on $X_i = x$, two compliance groups ($n$-never takers and $n$-compliers) have identified means of $Y_i(n)$, which the restriction in (14) implies obey the following relationships:

$$\mu_n^{nnt}(x) = \theta_n^0 + x'\theta_n^x + \gamma_n^h \bar{\lambda}_h^{nnt}(x) + \gamma_n^c \bar{\lambda}_c^{nnt}(x),$$

$$\mu_n^{nc}(x) = \theta_n^0 + x'\theta_n^x + \gamma_n^h \bar{\lambda}_h^{nc}(x) + \gamma_n^c \bar{\lambda}_c^{nc}(x),$$

where $nnt$ refers to $n$-never takers, $nc$ refers to $n$-compliers, and $\mu_n^g(x)$ is the mean of $Y_i(n)$ for the relevant compliance group $g$. The $\bar{\lambda}_d^g(x)$ terms are means of $v_{id}$ for each compliance group. These terms are similar to the control functions in (15), giving mean unobservables conditional on compliance group rather than program choice. We show in Appendix G that like the control function terms, the $\bar{\lambda}_d^g(x)$ are functions of parameters from the multinomial Probit model and are therefore identified.

Differencing these equations yields

$$\mu_n^{nnt}(x) - \mu_n^{nc}(x) = \gamma_n^h \left(\bar{\lambda}_h^{nnt}(x) - \bar{\lambda}_h^{nc}(x)\right) + \gamma_n^c \left(\bar{\lambda}_c^{nnt}(x) - \bar{\lambda}_c^{nc}(x)\right). \tag{16}$$

From this we see that the selection coefficients parameterize the gap in mean potential outcomes between compliance groups. With two available values of $X_i$, equation (16) can be evaluated twice, yielding two equations in the two unknown selection coefficients $\gamma_h^n$ and $\gamma_c^n$. When $X_i$ shifts the mean utilities for Head Start and competing programs in an non-degenerate fashion, these two equations can be solved for the selection coefficients. Similar arguments show identification of selection coefficients and means for $Y_i(h)$ and $Y_i(c)$. Appendix G formalizes the identification proof

[13] Although it is possible to extend the model to allow for upon higher-order polynomials in the selection terms (e.g., as in Dahl, 2002), reliable estimation in the HSIS sample would necessitate stronger instruments and larger samples than are presently available. Below we conduct some specification tests which indicate that (15) provides a reasonable fit to the data.

and provides explicit formulas in the case where a single binary covariate is available. Note that this identification argument has the flavor of a "difference in differences" strategy – the difference between compliance group means is compared across covariate groups to infer the selection coefficients.

Three important observations follow from this argument. First, identification of the selection coefficients holds even if $X_i$ is a fully saturated vector of exclusive and exhaustive indicator variables. Second, unlike the models justifying interacted IV approaches, (14) does not impose restrictions on average treatment effects across subgroups. The coefficients $\theta_d^x$ can vary in an unrestricted way, which allows for arbitrary treatment effect heterogeneity with respect to $X_i$. Finally, additive separability of the potential outcomes in observables and unobservables is essential for identification. If the selection coefficients in (16) were assumed to depend on $x$, there would be two unknowns for every value of $x$, and point 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 across observable subgroups to infer the parameters governing the selection process.

### Estimation

We estimate the model in two steps. First, we estimate the parameters of the Probit choice model via simulated maximum likelihood. The choice probabilities are efficiently evaluated using the Geweke-Hajivassiliou-Keane (GHK) simulator (Geweke, 1989; Hajivassiliou and McFadden, 1998; Keane, 1994). Second, we use the parameters of the choice model to form control function estimates $\left(\hat{\lambda}_{dh}\left(X_i, Z_i\right), \hat{\lambda}_{dc}\left(X_i, Z_i\right)\right)$, which are included as regressors in least squares estimation of (15).

When estimating the model we renorm the covariate vector $X_i$ to have unconditional mean zero so that the coefficients $\theta_d^0$ can be interpreted as average potential outcomes. Hence, the intercept differences $\theta_h^0 - \theta_n^0$ and $\theta_h^0 - \theta_c^0$ can be read as average treatment effects of Head Start relative to no preschool and other preschools. To increase precision, we also estimate models that restrict the coefficients $\left\{\theta_d^0, \theta_d^x, \gamma_d^h, \gamma_d^c\right\}_{d\in\{h,c,n\}}$ across program alternatives. Our preferred specifications restrict the degree of treatment effect heterogeneity present in the model by forcing some of these coefficients to be equal across alternatives $d$ – i.e., to effect an equal location shift in all three potential outcomes. We find that these restrictions fit the data well.

## 8 Structured estimates

### Parameter estimates

Table 8 reports estimates of the choice model. Column (1) shows the coefficients governing the mean utility of enrollment in Head Start. As expected, an offer to participate in Head Start substantially raises the implied utility of program enrollment. Moreover, the effects of an offer are greater at high-quality centers and especially at centers offering transportation services. Offers are less influential for poor households. We strongly reject the null hypothesis that the program offer interaction effects in the Head Start utility equation are insignificant. Because the main effects of the covariates $X_i^1$ were not randomly assigned, we cannot interpret them causally. However, some interesting patterns are present here as well. For example, households are less likely to participate in Head Start in the absence of an offer at sites with good transportation services.

Column (4) reports the parameters governing the correlation in unmeasured tastes for Head Start and competing programs. On average, the correlation is significantly positive, indicating that households view preschool alternatives as more similar to each other than to home care. This finding indicates that the IIA condition underlying logit-based choice models is empirically violated. There is some evidence of heterogeneity in the correlation based upon mother's education but we cannot reject the joint null hypothesis that the correlation is constant across covariate groups.

Table 9 reports second-step estimates of the parameters in (15). Column (1) omits all controls and simply reports naive differences in mean test scores across groups (the omitted category is home care). Head Start students achieve mean test scores roughly 0.2 standard deviations higher than students receiving 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, Column (2) 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). Unlike conventional value-added models, the controls are fully interacted with program alternative, making this a trichotomous generalization of the selection on observables framework studied by Oaxaca (1973) and Kline (2011). 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 fall significantly once controls are added.

Column (3) adds control functions adjusting for selection on unobservables. To account for uncertainty in the estimated control functions, inference for the two-step models is conducted via the nonparametric bootstrap, clustered by experimental site. Unlike the specifications in previous columns, identification of these control function terms relies on the experimental variation in offer assignment. The control function terms are jointly significant ($p$-value = 0.014), indicating a formal rejection of the selection on observables assumptions underlying value added specifications. Adjusting for selection on unobservables raises the estimated average impacts of Head Start and other preschools dramatically. However, the estimates are also very imprecise. Imprecision in average treatment effects is not surprising given that non-parametric identification of such quantities would require a large support assumption on the instrument (Heckman, 1990), which does not hold in our setting. More troubling is that many of the control function coefficients are imprecise despite being jointly significant, a sign that the control functions remain highly collinear despite the use of several program offer interactions.

To improve precision, we consider a variety of additional restrictions. Column (4) restricts a subset of the covariates to have common coefficients across program alternatives.[14] This improves

[14]Specifically, this restriction imposes that the quadratic and cubic terms in baseline score along with all covariates

the precision of the average treatment effect estimates along with some of the coefficients governing selection. Column (5) restricts the selection correction coefficients to be equal in the Head Start and competing preschool alternatives (i.e. $\gamma_h^h = \gamma_c^h$ and $\gamma_h^c = \gamma_c^c$), which is a natural restriction given that these preschools likely provide similar services. Finally, column (6) restricts the average treatment effect of Head Start to equal that of competing preschools (i.e. $\theta_h^0 = \theta_c^0$). None of these restrictions is rejected ($p$-values $\geq 0.539$), which bolsters our presumption that Head Start and competing preschools in fact provide similar educational services.

It is worth noting that even our most heavily constrained model reported in column (6), which will be our preferred specification, is still quite flexible, allowing for treatment effect heterogeneity with respect to baseline score and for selection into preschool based upon levels and gains. We find evidence for both sorts of selection in the data. Estimates of $\gamma_h^h$ are negative and statistically significant in all specifications. In other words, children from households with stronger tastes for Head Start have lower scores when attending Head Start. Our estimates of $\gamma_n^h$ are always statistically insignificant and usually close to zero. The difference $\gamma_h^h - \gamma_n^h$ is therefore negative, meaning that children who are more likely to attend Head Start receive smaller achievement benefits when shifted from home care to Head Start. This is inconsistent with Roy (1951)-style selection on test score gains, and suggests large benefits for children that are unlikely to attend the program.[15] This "reverse-Roy" pattern could reflect access issues (e.g. disadvantaged households living far from public transportation) or a lack of information about Head Start on the part of some households (as in the model of Appendix C).

In contrast, the estimated difference between $\gamma_c^c$ and $\gamma_n^c$ is always positive, though these coefficients are imprecise. This suggests that there may be positive selection on gains into other preschool programs. We reject the hypothesis of no selection on levels ($\gamma_k^d = 0\ \forall(k,d)$) in all specifications, and the hypothesis of no selection on gains ($\gamma_d^k = \gamma_j^k$ for $d \neq j,\ k \in \{h, c\}$) is rejected at the 10-percent level in our most precise specification.

### Model fit

Table 10 provides a specification test for our preferred restricted model by comparing mean potential outcomes for different compliance groups implied by the model to nonparametric IV estimates, wherever they exist. This Table also uses the model to impute the missing potential outcome means from Table 7. Reassuringly, the IV and structural estimates line up closely: the only discrepancies arise in the estimation of mean potential outcomes at competing preschools, and these discrepancies are small. The hypothesis that the fully-restricted structural model matches all moments is not rejected at conventional levels ($p$-value = 0.13).

Appendix Figure A1 provides an alternative check of the model's fit that more closely parallels the identification argument given in Section 7. To construct the Figure we computed values of

in $X_i^2$ besides race to have common coefficients in each alternative. We allow the coefficients on race dummies, the linear term on baseline score, and the elements of $X_i^1$ to differ across alternatives.

[15] Walters (2014) finds a similar pattern of selection in the context of charter schools.

$\bar{\lambda}_d^g(X_i)$, which give mean selection errors by compliance group, for each observation's value of $X_i$. For the three potential outcome equations we then computed gaps in mean selection errors for the two relevant compliance groups, and split the sample into nine cells defined by interactions of terciles of these two gaps (see Appendix G). In each cell we compute nonparametric differences in mean potential outcomes using the methods described in Appendix D. We then plot differences in potential outcomes against differences in mean selection errors, along with lines indicating the selection coefficients from the fully restricted model in Table 9. The model's predictions mirror the corresponding patterns in potential outcomes across compliance groups, though the estimates here are imprecise in some cases.

### SubLATEs

Table 11 reports implied treatment effects for each of our selection-corrected models. Identification of average treatment effects relies on parametric extrapolation beyond the population of program compliers, which leads to substantial imprecision in the point estimates. The model generates more precise estimates of the subLATEs, which we compute by integrating over the relevant regions of $X_i$, $v_{ih}$ and $v_{ic}$.

The first row of Table 11 uses the model parameters to compute the pooled $LATE_h$, which is nonparametrically identified by the experiment. Reassuringly, the model estimates line up closely with the nonparametric estimate obtained via IV. The remaining rows report estimates of effects relative to specific care alternatives. 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 0.35 standard deviations. By contrast, estimates of $LATE_{ch}$, though somewhat more variable across specifications, never differ significantly from zero. If anything, the $LATE_{ch}$ estimates suggest that Head Start is slightly more effective at boosting test scores than competing preschools.

It is worth comparing these 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 small effects for compliers drawn from other preschools, though their point estimate of $LATE_{nh}$ is somewhat smaller than ours (0.21 vs. 0.35). This difference reflects a combination of different test score outcomes (Feller et al. look only at PPVT scores) and different modeling assumptions. Specifically, their approach exploits a parametric prior over model parameters, restrictions on effect heterogeneity across subgroups, and a normality assumption on potential outcomes within each compliance group. By contrast, we consider a parametric choice model in conjunction with semi-parametric restrictions on the unselected distribution of potential outcomes. Since neither estimation approach nests the other, it is reassuring that these two approaches produce qualitatively similar findings.

# 9 Policy Simulations

## Structural reforms

The "reverse Roy" pattern revealed by the estimates in Table 10 suggests large potential effects of policies that target children who are currently unlikely to attend Head Start. Motivated by this finding, we next predict the social benefits of a reform that expands Head Start by making it more attractive rather than extending offers to additional households. We model this reform as an improvement in the structural program feature $f$, as described in Section 5. This can be viewed as a policy experiment that increases Head Start transportation services or outreach efforts to target children who would otherwise decline offers. We use our structured estimates to compute marginal treatment effects for these alternative subgroups of compliers, treating changes in $f$ as changes to the intercept $\psi_h^0$ of the Head Start utility in (13).

Figure 2 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 black vertical line indicating the current rate ($f = 0$). The right-hand axis measures $\overrightarrow{S}_c$, the share of marginal students drawn from other preschools, which is plotted in green. The left-hand axis measures test score effects. Red and yellow curves plot marginal treatment effects for subgroups of marginal students drawn from home care and other preschools, while the blue curve plots $MTE_h$, a weighted average of alternative-specific effects.

The figure shows that Head Start's marginal treatment effect increases with the scale of the program. 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, who experience larger score gains. This results in sharply increasing marginal treatment effects for Head Start relative to no preschool. Predicted effects for children drawn from other preschools are negligible for all values of $f$, which is unsurprising since we have imposed a zero average treatment effect of $h$ relative to $c$. Interestingly, the model predicts increasing crowdout of other preschools as Head Start expands, indicated by a rising value of $\overrightarrow{S}_c$. The overall marginal treatment effect therefore combines two offsetting forces: increasing gains for children drawn from home care, and an increasing share of children drawn from other preschools for whom gains are negligible. This leads to a modestly increasing schedule of $MTE_h$.

To investigate the consequences of this pattern for the social return to Head Start, Figure 3 plots $MVPF_f$, the marginal value of public funds for structural reforms. This Figure relies on the same parameter calibrations as Table 6. Calculations of $MVPF_f$ must account for the fact that changes in structural program features may increase the direct costs of the program. The term $\phi_h'(f)\left(\partial \ln P\left(D_i = h\right)/\partial f\right)^{-1}$ in equation (12) captures this effect. This term can be written $\eta \cdot \phi_h$, where $\eta = d\ln\phi_h/d\ln P(D_i = h)$ is 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 $\eta$. We start with the extreme case where $\eta = 0$, which allows us to characterize costs and benefits associated with reforms that draw in children on the margin without changing the per-capita cost of the program. We then consider how the cost-benefit calculus changes when $\eta > 0$.

As in our basic cost/benefit analysis, the results in Figure 3 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 includes the attenuated benefits generated by increasing crowdout while effectively ignoring the attenuated costs. The blue curve accounts for public savings by setting $\phi_c$ equal to our preferred value of $0.75\phi_h$. This generates an upward shift and steepens the $MVPF_f$ schedule, indicating higher social returns that increase more quickly with program size. The implied marginal value of public funds at $f = 0$ is above 2. This is larger than 1.84, the value of $MVPF_\delta$ reported in Table 6, which implies larger social returns for expansions that attract new households than for expansions that raise the offer rate.

Finally, the green line in Figure 3 shows $MVPF_f$ when $\phi_c = 0.75\phi_h$ and $\eta = 0.5$.[16] This scenario implies sharply rising marginal costs of Head Start provision: an increase in $f$ that doubles enrollment 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 $\eta$ 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, as larger expansions draw in households with weaker tastes for preschool with large potential gains. The social returns to attracting such students will be substantial unless program costs increase rapidly with enrollment. These findings imply that structural reforms targeting children who are currently unlikely to attend Head Start and children who 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).

### Rationing of competing preschools

In the cost-benefit analysis in Section 6 we assumed that seats at competing preschools are not rationed. This implies that when the Head Start offer probability $\delta$ increases competing preschools contract in size rather than admitting more students. While this assumption is reasonable for some programs such as universal state preschool, other public programs may face relatively fixed

[16] 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.

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 competing preschools rather than generating cost savings in these programs. Our structured estimates allow us to assess the sensitivity of our cost/benefit results to the possibility of rationing in competing programs.

In Appendix B we adapt the model of Section 5 to a setting with rationing in other preschools. We assume that competing programs adjust their offer rates so that total enrollment in option $c$ does not change when $\delta$ increases. Furthermore, we assume that children drawn into seats vacated by new Head Start enrollees would otherwise receive home care. In this case, the $MVPF$ for an increase in $\delta$ becomes:

$$MVPF_{\delta} = \frac{(1-\tau)p\,(LATE_h + S_c LATE_{nc})}{\phi_h - \tau p\,(LATE_h + S_c LATE_{nc})}, \tag{17}$$

where $LATE_{nc}$ measures the average effect of competing preschools relative to home care for children who comply with offers from alternative $c$. The denominator of (17) does not include the cost savings term in (9) because rationed competing programs expand to fill seats vacated by Head Start enrollees. The numerator includes the extra term $(1-\tau)pS_c LATE_{nc}$, which measures the earnings gain for children who fill the vacated seats.

We compute social rates of return under three alternative assumptions about $LATE_{nc}$. First, we assume that $LATE_{nc} = 0$. Second, we assume that 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 structured estimates to compute a prediction for $LATE_{nc}$. Specifically, we add the offer coefficient $\psi_h^z$ to the other center utility $U_i(c)$ and then compute average treatment effects for students induced to switch from $n$ to $c$ by this change.

Appendix Table A5 shows the results of this analysis. Setting $LATE_{nc} = 0$ yields an $MVPF$ 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 ratio of 2.44. The model predicts that $LATE_{nc} = 0.446$, which produces a ratio of 2.85. These results suggest that our cost/benefit results are robust to the possibility of rationing in competing programs. Under plausible assumptions about the effects of competing programs relative to home care, accounting for the benefits generated by vacated seats in these programs produces social returns larger than those displayed in panel B of Table 6.

# 10 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 informative for the debate over calls for universal preschool, which might reach high return households. One would need adequate projections of the cost of providing such services in order to assess the return to such proposals.

It is important to note some 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 are available on enrollment in competing programs.

## 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. Abadie, A. (2002). "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variables Models." Journal of the American Statistical Association 97(457).

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

4. Acemoglu, D. (2001). "Good jobs versus bad jobs." Journal of Labor Economics, 19(1), 1-21.

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

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

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

8. Behaghel, L., Crepon, B., and Gurgand, M. (2013). "Robustness of the Encouragement Design in a Two-Treatment Randomized Controlled Trial." IZA Discussion Paper no. 7447.

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

10. 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.

11. Carneiro, P., and Lee, S. (2009). "Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality." Journal of Econometrics 149(2).

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. Cascio, E., and Staiger, D. (2012). "Knowledge, Tests, and Fadeout in Educational Interventions." NBER Working Paper no. 18038.

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., & Saez, E. (2014). "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. Cunha, F., Heckman, J., and Schennach, S. (2010). "Estimating the Technology of Cognitive and Non-cognitive Skill Formation." Econometrica 78(3).

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

22. 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.

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

24. Dahl, G. (2002). "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets." Econometrica 70.

25. DasGupta, A. (2008). Asymptotic theory of statistics and probability. Springer Science & Business Media.

26. Deaton, A. (1985). "Panel Data From Time Series of Cross-sections." Journal of Econometrics 30(1).

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

28. Devereux, P. J. (2007). "Improved Errors-in-variables Estimators for Grouped Data." Journal of Business and Economic Statistics 25(3).

29. Doyle, J. (2007). "Child Protection and Child Outcomes: Measuring the Effects of Foster Care." American Economic Review 97(5).

30. Doyle, J. (2008). "Child Protection and Adult Crime: Using Investigator Assignment to Estimate Causal Effects of Foster Care." Journal of Political Economy 116(4).

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

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

33. 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.

34. 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).

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

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

37. 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.

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

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

40. Goldberger, A., and Olkin, I. (1971). "A Minimum-distance Interpretation of Limited-information Estimation." Econometrica 39(3).

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

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

43. Havnes, T., and Mogstad, M. (2011). "No Child Left Behind: Subsidized Child Care and Children's Long-run Outcomes." American Economic Journal: Economic Policy 3(2).

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

45. Heckman, J. (1990). "Varieties of Selection Bias." American Economic Review 80(2).

46. Heckman, J., and Smith, J. (1999). "Evaluating the Welfare State." In Steiner Strom, ed., *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium*. Cambridge, UK: Cambridge University Press.

47. 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).

48. Heckman, J., and Vytlacil, E. (2001). "Policy-relevant Treatment Effects." American Economic Review 91(2).

49. Heckman, J., and Vytlacil, E. (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." Econometrica 73.

50. 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).

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

52. 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).

53. 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).

54. 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).

55. Heckman, J., Urzua, S., and Vytlacil, E. (2006). "Understanding Instrumental Variables in Models with Essential Heterogeneity." The Review of Economics and Statistics 88(3).

56. 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.

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

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

59. Imbens, G., and Rubin, D. (1997). "Estimating Outcome Distributions for Compliers in Instrumental Variables Models." The Review of Economic Studies 64(4).

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

61. Kahneman, D., Wakker, P., and Sarin, R. (1997). "Back to Bentham? Explorations of Experienced Utility." Quarterly Journal of Economics 112(2).

62. 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).

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

64. Kirkeboen, L., Leuven, E., and Mogstad, M. (2014). "Field of Study, Earnings, and Self-Selection." Mimeo, University of Chicago.

65. 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`.

66. Kline, P. (2011). "Oaxaca-Blinder as a Reweighting Estimator." American Economic Review: Papers and Proceedings 101(3).

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

68. Krueger, A. (1999). "Experimental Estimates of Education Production Functions." Quarterly Journal of Economics 114 (2).

69. Lafontaine, F., and White, K. J. (1986). "Obtaining any Wald Statistic you Want." Economics Letters 21(1), 35-40.

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

71. 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).

72. 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).

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

74. Mayshar, J. (1990). "On measures of excess burden and their application." Journal of Public Economics, 43(3), 263-289.

75. 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).

76. Mogstad, M., and Wiswall, M. (2010). "Testing the Quantity-Quality Model of Fertility: Linearity, Marginal Effects, and Total Effects." NYU Working Paper.

77. Murnane, Willett and Levy (1995). "The Growing Importance of Cognitive Skills in Wage Determination." The Review of Economics and Statistics 77(2).

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

79. Oaxaca, R. (1973). "Male-Female Wage Differentials in Urban Labor Markets." International Economic Review 14(3).

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

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

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

83. Samuelson, P. (1954). "The Pure Theory of Public Expenditure." The Review of Economics and Statistics 36(4).

84. Saez, E., Matsaganis, M., and Tsakloglou, P. (2012). "Earnings Determination and Taxes: Evidence From a Cohort-Based Payroll Tax Reform in Greece." The Quarterly Journal of Economics 127(1).

85. Saez, E., Slemrod, J., and Giertz, S. (2012). "The Elasticity of Taxable Income With Respect to Marginal Tax Rates: A Critical Review." Journal of Economic Literature 50(1).

86. 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.

87. Sojourner, A. (2009). "Inference on Peer Effects with Missing Peer Data: Evidence from Project STAR." Working Paper.

88. Solon, G. (2002). "Cross-country Differences in Intergenerational Mobility." Journal of Economic Perspectives 16(3).

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

90. Tallis, G. (1961). "The Moment Generating Function of the Truncated Multi-normal Distribution." Journal of the Royal Statistical Society 23(1).

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

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

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

94. 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 .

95. 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.

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

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

# 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: DHHS (2010) shows estimates for each individual test, Walters (forthcoming) 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 (DHHS 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;" "Head Start;" 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 (DHHS 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 DHHS (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: Model

This appendix derives the conditions for optimal program scale and program features in equations (9), (12) and (17).

## Optimal program scale

First, consider the government's optimal choice of $\delta$. From (6) and (7), the first-order condition for the optimal value of $\delta$ is given by

$$(1-\tau)\, p \frac{\partial E\left[Y_i\right]}{\partial \delta} = m\left(\phi_h \frac{\partial P\left(D_i = h\right)}{\partial \delta} + \phi_c \frac{\partial P\left(D_i = c\right)}{\partial \delta} - \tau p \frac{\partial}{\partial \delta} E\left[Y_i\right]\right),$$

where $m$ is a Lagrange multiplier. Differentiating with respect to $\tau$ and rearranging reveals that $m = 1$.

Random assignment of offers implies that

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

Therefore, we have

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

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

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

which is equation (8).

Next, we can write:

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

so that

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

$$= E\left[1\left\{D_i(1) = h, D_i(0) \neq h\right\}\right]$$

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

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

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

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

$$= -P\left(D_i(1) = h, D_i(0) = c\right).$$

Plugging the derivatives into the government's first-order condition, we have

$$\begin{aligned}(1-\tau)pLATE_h P(D_i(1) = h, D_i(0) \neq h) = {} & \phi_h P\left(D_i(1) = h, D_i(0) \neq h\right) \\ & - \phi_c P(D_i(1) = h, D_i(0) = c) \\ & - \tau pLATE_h P(D_i(1) = h, D_i(0) \neq h).\end{aligned}$$

Dividing both sides of this equation by $P\left(D_i(1) = h, D_i(0) \neq h\right)$ yields the result in (9).

## Optimal program features

Next, consider the optimal value of the program feature $f$. The first-order condition is:

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

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 $h$:

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

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

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

The corresponding effects on choice probabilities are

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

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

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 government's first-order condition and dividing by the total density of marginal compliers, we obtain

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

which is equation (12).

### Rationing of competing preschools

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 $\delta_h$ and $\delta_c$ 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 $\delta_c$ so that $dP(D_i = c)/d\delta_h = 0$. Totally differentiating equation (18) with respect to $\delta_h$ 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, $\delta_c$ 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

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

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

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

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

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

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

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

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

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

The government's first-order condition is

$$p(1-\tau)\cdot\frac{dE\left[Y_i\right]}{d\delta_h} = \phi_h\cdot\frac{dP\left(D_i = h\right)}{d\delta_h} - \tau p\cdot\frac{dE\left[Y_i\right]}{d\delta_h}$$

Since $\delta_c$ 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 first-order condition is

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

Dividing the left-hand side of this equation by the right-hand side gives (17).

# Appendix C: Optimal Head Start Provision Under a Utilitarian Social Objective

In this Appendix, we consider a more complex utilitarian planning problem and show that it yields results similar to the simple paternalistic planning problem described in the text. The key insight of the model is that preschool enrollment decisions are made by imperfectly altruistic parents who may choose for their children different educational options than the children would have (retrospectively) chosen for themselves. Since today's children are tomorrow's adults, this can lead to chronic underinvestment in human capital, which provides a potential rationale for subsidized provision of preschool. We start by providing obtaining an expression for the marginal value of public funds implied by the model, and then implement a number of simplifications that yield a lower bound MVPF coinciding with expression (10) in the text.

Each parent chooses a preschool enrollment option $d \in \{h, c, n\}$ for their child. Adults have preferences over (lifetime) consumption $q$ and leisure $l$ given by the utility function $u(q, l)$, which we assume is twice differentiable and concave increasing in both its arguments. The lifetime budget constraint of a parent with human capital endowment $y$ can be written:

$$q = (1 - \tau) py (T - l) - \kappa 1 \{d = c\},$$

where $p$ is the market price of human capital, $T$ is a time endowment that we assume is common across parents, and $\kappa$ is the tuition charged by competing preschool programs.

In addition to consumption and leisure, parents also value the expected welfare of their children. Changing notation slightly, each parent $i$ has indirect utility over the enrollment options of their child given by the function $U_i^d(l, y, z)$ where $z$ denotes whether or not a Head Start offer has been received. The alternative specific indirect utilities can be defined recursively as:

$$U_i^d(l, y, z) = u((1 - \tau) py (T - l) - \kappa 1 \{d = c\}, l) + \chi 1 \{d = h\} z + \lambda E_\omega V_i (Y_i(d), Z_i) + \varepsilon_i(d), \quad (19)$$

where $\chi$ is the utility value of a Head Start offer, $\{Y_i(h), Y_i(c), Y_i(n)\}$ are the potential test scores of parent $i$'s child, $\{\varepsilon_i(h), \varepsilon_i(c), \varepsilon_i(n)\}$ are parent $i$'s idiosyncratic expectational errors that are drawn from a continuously differentiable joint distribution function $F_\varepsilon(.)$, and

$$V_i(y, z) = \max_{d \in \{h,c,n\}} \left\{ \max_l U_i^d(l, y, z) \right\}$$

is the parent's value function. Under weak regularity conditions, the value function $V_i(l, y, z)$ will exist whenever $\lambda < 1$.

The first term in (19) gives the utility associated with the parent's labor supply choices. Labor supply decisions obey the static first order condition:

$$\frac{u_l((1 - \tau) py (T - l) - \kappa 1 \{d = c\}, l)}{u_q((1 - \tau) py (T - l) - \kappa 1 \{d = c\}, l)} = (1 - \tau) py$$

which implies the parent's leisure can be written as a function $l^* = l((1-\tau)y, d)$ with corresponding consumption choice $q^* = q((1-\tau)y, d)$ and indirect utility $u^*((1-\tau)y, d) = u(q^*, l^*)$.

The second term in (19) gives the effect of a Head Start offer on the attractiveness to parents of Head Start. Equivalently, we could have associated the absence of an offer with a hassle cost of enrolling in another Head Start lottery. The third term captures the altruism parents feel towards their children. The parameter $\lambda \in [0, 1)$ governs the weight placed on the utility the child will realize as an adult. Parents cannot forecast their child's potential test scores with certainty but rather form expectations based upon their information set $\omega_i$. The term $E_\omega V_i(Y_i(d), Z_i) = E[V_i(Y_i(d), Z_i)|\omega_i]$ gives the rational component of household expectation formation that coincides with the true conditional expectation. However, household expectations are not entirely rational because households with identical information diverge in their expectations according to the $\varepsilon_i(d)$ terms. Note that without these expectational errors households would engage in Roy-selection on gains of the sort that we reject in our later empirical analysis.

Consider the consequences of this preference structure in a setting with many identical generations.[17] Incomplete altruism ($\lambda < 1$) yields the possibility that some children, when they become adults, would be willing to have paid their parents to invest more in their human capital. This reveals the fundamental market failure present in the model: adults cannot pay to remedy human capital deficits that arise in early childhood.

We follow Acemoglu (2001) in abstracting from transitional dynamics and supposing the planner's problem is to simply choose among steady states. In a steady state equilibrium, the test score distributions of parents and children will be identical, therefore average household utility can be written $E[V_i(Y_i, Z_i)]$. It follows from (19) and the assumption that generations have identical distributions of heterogeneity that:

$$E[V_i(Y_i, Z_i)] = \frac{E[u^*((1-\tau)Y_i, D_i)] + \chi \cdot P(D_i(1) = h)\delta + E[\varepsilon_i]}{1-\lambda},$$

where $\varepsilon_i \equiv \sum_{d\in\{h,c,n\}} 1[D_i = d] \cdot \varepsilon_i(d)$. Since $E[\varepsilon_i]$ is the component of utility resulting from expectational errors, it is natural to disregard it when solving a utilitarian planning problem and to focus on maximizing "experienced utility" (Kahneman, Wakker, and Sarin, 1997). Accordingly, we write the planner's objective as:

$$W = \underbrace{E[u^*((1-\tau)Y_i, D_i)]}_{\text{Consumption/Leisure}} + \underbrace{\chi \cdot P(D_i(1) = h)\delta}_{\text{Offer Utility}}, \tag{20}$$

where we have ignored the component of the household's "decision utility" associated with the expectational errors $\varepsilon_i$. In words, this criterion is simply the average utility from consumption and leisure in the economy plus the psychic benefits enjoyed by Head Start participants with lottery offers.

[17] By identical, we mean that successive cohorts have identical joint distributions of potential test scores, skill prices, and idiosyncratic tastes.

The first order condition for maximization of (20) subject to (7) can be written:

$$\begin{aligned}\underbrace{LATE_h^u}_{\text{Utility Gain}} + \underbrace{\chi}_{\text{Complier Rent}} &+ \underbrace{\chi\frac{P\left(D_i\left(0\right)=h\right)}{P\left(D_i\left(1\right)=h\right)-P\left(D_i\left(0\right)=h\right)}}_{\text{Rent to Always Takers}} \\ &= m\left[\underbrace{\phi_h}_{\text{Provision Cost}} - \underbrace{\phi_c S_c}_{\text{Public Savings}} - \underbrace{\tau p LATE_h}_{\text{Added Revenue}}\right].\end{aligned} \tag{21}$$

where $LATE_h^u \equiv E\left[u^*\left(\left(1-\tau\right)Y_i\left(h\right),h\right) - u^*\left(\left(1-\tau\right)Y_i\left(D_i\left(0\right)\right),D_i\left(0\right)\right)|D_i\left(1\right)=h, D_i\left(0\right)\neq h\right]$ is the LATE on consumption utility of compliers and $m$ is a Lagrange multiplier. The left hand side of (21) gives the net effect on welfare of enrolling another child in Head Start. An average program complier experiences a consumption utility gain of $LATE_h^u$. Compliers also receive a non-pecuniary utility gain of $\chi$ associated with the Head Start offer. To attract a program complier via random lottery we must in expectation also make offers to a fraction $\frac{P(D_i(0)=h)}{P(D_i(1)=h)-P(D_i(0)=h)}$ of Head Start "always takers", who would have enrolled without the offer but now save the disutility $\chi$ of otherwise having to shop around centers to enroll in Head Start.

The right hand side of (21) gives the cost to government of adding another Head Start slot, which is the same as the corresponding expression in (9) except that there is now a non-trivial Lagrange multiplier $m$ translating dollars into marginal utility. Hence, the marginal value of public funds associated with a change in the offer probability $\delta$ can be written:

$$MVPF_\delta = \frac{LATE_h^u + \chi\frac{P(D_i(1)=h)}{P(D_i(1)=h)-P(D_i(0)=h)}}{m\left(\phi_h - \phi_c S_c - \tau p LATE_h\right)}.$$

We now seek to simplify this expression and convert it into an identifiable lower bound.

## A Lower Bound

To obtain a lower bound, we first utilize the fact that empirical studies (e.g., Saez, Matsaganis, and Tsakloglou, 2012) often find *uncompensated* lifetime labor supply elasticities near zero, suggesting that $\frac{\partial}{\partial y} l\left(\left(1-\tau\right)y, d\right) \approx 0$. We therefore treat $l^*$ as a constant.

With this simplification, it is straightforward to deduce the multiplier $m$ by taking the planner's first order condition with respect to $\tau$ and rearranging, which yields:

$$m = \left(T - l^*\right)\frac{E\left[Y_i u_q\left(q^*, l^*\right)\right]}{E\left[Y_i\right]}.$$

The multiplier $m$ is the human capital weighted average marginal utility of consumption times the amount worked. Since human capital and the marginal utility of consumption are negatively correlated, we expect:

$$m \leq \left(T - l^*\right) E\left[u_q\left(q^*, l^*\right)\right]. \tag{22}$$

That is, we expect the equally weighted average marginal utility of consumption to exceed the human capital weighted version.

Next consider the term $LATE_h^u$, which we can decompose as follows:

$$\begin{aligned}
LATE_h^u &= E\left[u^*\left(\left(1-\tau\right)Y_i\left(h\right),h\right)-u^*\left(\left(1-\tau\right)Y_i\left(D_i\left(0\right)\right),n\right)|D_i\left(1\right)=h,D_i\left(0\right)\neq h\right] \\
&\quad +E\left[u^*\left(\left(1-\tau\right)Y_i\left(D_i\left(0\right)\right),n\right)-u^*\left(\left(1-\tau\right)Y_i\left(D_i\left(0\right)\right),D_i\left(0\right)\right)|D_i\left(1\right)=h,D_i\left(0\right)\neq h\right]
\end{aligned}$$

This second term evaluates to

$$S_cE\left[u\left(\left(1-\tau\right)pY_i\left(c\right)\left(T-l^*\right),l^*\right)-u\left(\left(1-\tau\right)pY_i\left(c\right)\left(T-l^*\right)-\kappa,l^*\right)|D_i\left(1\right)=h,D_i\left(0\right)=c\right]>0,$$

which is the average utility increase associated with the tuition savings from reducing enrollment in other programs. Since this term must be positive, a lower bound is given by the first term, which we consider from here on.

Note that by a mean value expansion, for any household $i$, we have

$$\begin{aligned}
u^*\left(\left(1-\tau\right)Y_i\left(h\right),h\right)-u^*\left(\left(1-\tau\right)Y_i\left(D_i\left(0\right)\right),n\right) &= u\left(\left(1-\tau\right)pY_i\left(h\right)\left(T-l^*\right),l^*\right) \\
&\quad -u\left(\left(1-\tau\right)pY_i\left(D_i\left(0\right)\right)\left(T-l^*\right),l^*\right) \\
&= u_q\left(\left(1-\tau\right)p\bar{Y}_i\left(T-l^*\right),l^*\right) \\
&\quad \times\left(1-\tau\right)p\left[Y_i\left(h\right)-Y_i\left(D_i\left(0\right)\right)\right]\left(T-l^*\right)
\end{aligned}$$

for some $\bar{Y}_i\in\left(Y_i\left(D_i\left(0\right)\right),Y_i\left(h\right)\right)$. Hence, we have:

$$LATE_h^u\geq\left(1-\tau\right)p\left(T-l^*\right)E\left[u_q\left(\left(1-\tau\right)p\bar{Y}_i\left(T-l^*\right),l^*\right)\left(Y_i\left(h\right)-Y_i\left(D_i\left(0\right)\right)\right)|D_i\left(1\right)=h,D_i\left(0\right)\neq h\right].$$

Estimates from Bitler, Domina, and Hoynes (2015) and Table 7 of Walters (forthcoming) suggest that Head Start treatment effects are negatively correlated with human capital levels (i.e. that Head Start impacts are "compensatory"). Because the marginal utility of consumption is declining in earnings, this suggests that $Cov\left(u_q\left(\left(1-\tau\right)p\bar{Y}_i\left(T-l^*\right),l^*\right),\left(Y_i\left(h\right)-Y_i\left(D_i\left(0\right)\right)\right)\right)>0$. As a result, we have the additional bound:

$$LATE_h^u\geq\left(1-\tau\right)p\left(T-l^*\right)E\left[u_q\left(\left(1-\tau\right)p\bar{Y}_i\left(T-l^*\right),l^*\right)|D_i\left(1\right)=h,D_i\left(0\right)\neq h\right]LATE_h. \quad (23)$$

Combining the upper bound in (22) with the lower bound in (23), we arrive at a lower bound for $MVPF_\delta$:

$$\underline{MVPF_\delta}=\frac{\left(1-\tau\right)pLATE_h}{\phi_h-\phi_cS_c-\tau pLATE_h}\frac{E\left[u_q\left(\left(1-\tau\right)p\bar{Y}_i\left(T-l^*\right),l^*\right)|D_i\left(1\right)=h,D_i\left(0\right)\neq h\right]}{E\left[u_q\left(q^*,l^*\right)\right]}.$$

If the average marginal utility of consumption among compliers (evaluated at intermediate human capital levels $\bar{Y}_i$) equals the average marginal utility of consumption in the entire population, then this expression coincides exactly with the paternalistic MVPF listed in (10). In practice, compliers seem to be somewhat positively selected in terms of their average test score outcomes (see Table 10), which could make this ratio less than one if the curvature of utility is substantial over the income range being examined. However, since the differences in expected income between compliers and other groups are small, and compliers are roughly 70 percent of the study population, any deviation from a ratio of one is likely to be quantitatively unimportant in practice.

## Appendix D: Identification of Complier Characteristics

This appendix extends results from Abadie (2002) to show identification of characteristics and marginal potential outcome distributions for subpopulations of compliers drawn from other preschools and no preschool. Let $g(Y_i, X_i)$ 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

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

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

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

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

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

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

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

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 $\kappa_c$ is the population coefficient from an instrumental variables regression of $g(Y_i, X_i) \cdot 1\left\{D_i = c\right\}$ on $1\left\{D_i = c\right\}$, 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 Tables 4 and 5 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 Tables 7 and 10 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 E: Interacted Two-stage Least Squares

This Appendix investigates the use of the interacted two-stage least squares approach described in Section 7 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 4, 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 $1\{D_i=h\}$ and $1\{D_i=c\}$ 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

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

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

which can be written

$$\begin{aligned} LATE_{nh} = {} & \tfrac{S_c(0)S_c(1)}{S_c(0)-S_c(1)} \cdot (LATE_{ch}(1) - LATE_{ch}(0)) \\ & + (w_n(1) LATE_{nh}(1) + (1 - w_n(1)) LATE_{nh}(0)), \end{aligned} \tag{24}$$

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.

### Interacted 2SLS estimates

To empirically explore the interacted 2SLS strategy we generate additional instruments by interacting the Head Start offer indicator with the vector $X_i^1$ (introduced in equation 13) which consists of indicators for whether a child's center of random assignment offers transportation, above-median center quality, mother's education, age four, and an indicator for family income above the poverty line. Table A4 displays the resulting interacted 2SLS estimates. The first row in column (1) displays the 2SLS estimate of $LATE_h$ pooled across age cohorts in the first year of the experiment. This yields the same value of 0.247 found in column (9) of Table 2. The second row uses as additional instruments linear interactions of the $Z_i$ and $X_i^1$. Adding these instruments slightly improves precision but leads to a rejection of the model's overidentifying restrictions ($p$-value = 0.048). This may be caused either by heterogeneity in $S_c$ or the subLATEs.

Columns (2) and (3) of Table A3 treat Head Start and other preschools as separate endogenous variables. The resulting 2SLS estimates suggests that Head Start and competing preschools have roughly similar effects, but the instruments generate relatively weak independent variation in competing preschools: the Angrist and Pischke (2009) partial F-statistic for other preschools is 1.9, far short of the standard rule of thumb of 10.0. Nonetheless, the overidentifying restrictions are still rejected ($p = 0.038$), which suggests the presence of additional heterogeneity not explained by counterfactual fallbacks. Equation (24) shows that this rejection may be caused by heterogeneity in either or both of the subLATEs, which implies that at least one of the two preschool coefficients cannot be interpreted as a well-defined causal effect.

## Appendix F: Control Functions

This appendix derives the control function terms used in the two-step models in Section 7. For ease of notation, we rewrite the model in (13) as

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

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

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

Note that 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

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

where

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

Defining $\lambda_{dk}\left(X_i, Z_i\right) \equiv E\left[v_{ik} | X_i, Z_i, D_i = d\right]$, this implies that we can write

$$\begin{aligned} \lambda_{hh}(X_i, Z_i) &= -\Lambda\left(\psi_h(X_i), \frac{\psi_h(X_i,Z_i)-\psi_c(X_i)}{\sqrt{2(1-\rho(X_i))}}; \sqrt{\frac{1-\rho(X_i)}{2}}\right), \\ \lambda_{hc}(X_i, Z_i) &= -\Lambda\left(\psi_h(X_i), \frac{\psi_h(X_i,Z_i)-\psi_c(X_i)}{\sqrt{2(1-\rho(X_i))}}; \sqrt{\frac{1-\rho(X_i)}{2}}\right) \\ &\quad + \sqrt{2(1-\rho(X_i))} \cdot \Lambda\left(\frac{\psi_h(X_i,Z_i)-\psi_c(X_i)}{\sqrt{2(1-\rho(X_i))}}, \psi_h(X_i); \sqrt{\frac{1-\rho(X_i)}{2}}\right). \end{aligned}$$

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

$$
\begin{aligned}
\lambda_{ch}(X_i, Z_i) =&\ -\Lambda\left(\psi_c(X_i), \tfrac{\psi_c(X_i)-\psi_h(X_i,Z_i)}{\sqrt{2(1-\rho(X_i))}}; \sqrt{\tfrac{1-\rho(X_i)}{2}}\right) \\
&+ \sqrt{2\left(1-\rho(X_i)\right)} \cdot \Lambda\left(\tfrac{\psi_c(X_i)-\psi_h(X_i,Z_i)}{\sqrt{2(1-\rho(X_i))}}, \psi_c(X_i); \sqrt{\tfrac{1-\rho(X_i)}{2}}\right), \\
\lambda_{cc}(X_i, Z_i) =&\ -\Lambda\left(\psi_c(X_i), \tfrac{\psi_c(X_i)-\psi_h(X_i,Z_i)}{\sqrt{2(1-\rho(X_i))}}; \sqrt{\tfrac{1-\rho(X_i)}{2}}\right), \\
\lambda_{nh}(X_i, Z_i) =&\ \Lambda\left(-\psi_h(X_i, Z_i), -\psi_c(X_i); \rho(X_i)\right), \\
\lambda_{nc}(X_i, Z_i) =&\ \Lambda\left(-\psi_c(X_i), -\psi_h(X_i, Z_i); \rho(X_i)\right).
\end{aligned}
$$

## Appendix G: Identification of the Selection Model

### Setup and notation

This appendix proves identification of the selection model for the case of a saturated model with one binary covariate, $X_i \in \{0,1\}$. As noted in Section 4, households can be partitioned into five groups: $n$-compliers, $c$-compliers, $n$-never takers, $c$-never takers, and always takers. Let $G_i \in \{nc, cc, nnt, cnt, at\}$ indicate group membership for household $i$, and let

$$\mu_d^g(x) = E\left[Y_i(d)|G_i = g, X_i = x\right]$$

denote the mean of potential outcome $d$ for group $g$. By the arguments presented in Appendix D, we can nonparametrically identify the following six mean potential outcomes:

$$\mu_n^{nc}(x) = E\left[Y_i(n)|G_i = nc, X_i = x\right],$$

$$\mu_c^{cc}(x) = E\left[Y_i(c)|G_i = cc, X_i = x\right],$$

$$\mu_h^{c}(x) = E\left[Y_i(h)|G_i \in \{nc, cc\}, X_i = x\right],$$

$$\mu_n^{nnt}(x) = E\left[Y_i(n)|G_i = nnt, X_i = x\right],$$

$$\mu_c^{cnt}(x) = E\left[Y_i(c)|G_i = cnt, X_i = x\right],$$

$$\mu_h^{at}(x) = E\left[Y_i(h)|G_i = at, X_i = x\right].$$

We can also nonparametrically identify the population fractions of each of these five groups, $\omega^g(x) = P\left(G_i = g|X_i = x\right)$.

We are interested in estimating subLATEs, local average treatment effects for compliers drawn from particular counterfactual alternatives, such as

$$LATE_{nh}(x) = \mu_h^{nc}(x) - \mu_n^{nc}(x).$$

This parameter is not nonparametrically identified because $\mu_h^{nc}(x)$ is not nonparametrically identified. However, suppose we impose an additively separable structure on potential outcomes:

$$E\left[Y_i(d)|X_i, v_{ih}, v_{ic}\right] = \mu_d(X_i) + \gamma_d^h v_{ih} + \gamma_d^c v_{ic}. \tag{25}$$

We next show how to use this restriction to identify the selection parameters $\gamma_d^h$ and $\gamma_d^c$, the mean potential outcomes $\mu_d(x)$, and the subLATEs.

## Choice probabilities

Before proving identification, it will be useful to characterize the link between nonparametric choice probabilities and the multinomial Probit model. The choice model can be represented

$$U_i(h, X_i, Z_i) = \psi_h(X_i, Z_i) + v_{ih},$$

$$U_i(c, X_i) = \psi_c(X_i) + v_{ic},$$

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

With a single binary covariate, the 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 choice conditional probabilities

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

To prove identification, we will need to work with average selection errors for each compliance group. Consider the quantity

$$\bar{\lambda}_d^g(x) = E\left[v_{id}|G_i = g, X_i = x\right].$$

This is the mean of the selection error from the multinomial probit model for a compliance group conditional on $X_i$. To compute these terms, define the function

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

Tallis (1961) shows that if $U$ and $V$ have a bivariate standard normal distribution with correlation $\xi$, then

$$E\left[U|a_0 < U < a_1, b_0 < V < b_1\right] = \Lambda_0(a_0, a_1, b_0, b_1; \xi).$$

It follows that

$$
\begin{aligned}
\bar{\lambda}_h^{nc}(x) =&\Lambda_0\left(-\psi_h(x,1),-\psi_h(x,0),-\infty,-\psi_c(x);\rho(x)\right),\\
\bar{\lambda}_c^{nc}(x) =&\Lambda_0\left(-\infty,-\psi_c(x),-\psi_h(x,1),-\psi_h(x,0);\rho(x)\right),\\
\bar{\lambda}_h^{cc}(x) =&\Lambda_0\left(-\psi_c(x),\infty,\tfrac{\psi_h(x,0)-\psi_c(x)}{\sqrt{2(1-\rho(x))}},\tfrac{\psi_h(x,1)-\psi_c(x)}{\sqrt{2(1-\rho(x)}};\sqrt{\tfrac{1-\rho(x)}{2}}\right)\\
&-\sqrt{2(1-\rho(x))}\Lambda_0\left(\tfrac{\psi_h(x,0)-\psi_c(x)}{\sqrt{2(1-\rho(x))}},\tfrac{\psi_h(x,1)-\psi_c(x)}{\sqrt{2(1-\rho(x)}},-\psi_c(x),\infty;\sqrt{\tfrac{1-\rho(x)}{2}}\right),\\
\bar{\lambda}_c^{cc}(x) =&\Lambda_0\left(-\psi_c(x),\infty,\tfrac{\psi_h(x,0)-\psi_c(x)}{\sqrt{2(1-\rho(x))}},\tfrac{\psi_h(x,1)-\psi_c(x)}{\sqrt{2(1-\rho(x)}};\sqrt{\tfrac{1-\rho(x)}{2}}\right),\\
\bar{\lambda}_h^{nnt}(x) =&\Lambda_0\left(-\infty,-\psi_h(x,1),-\infty,-\psi_c(x);\rho(x)\right),\\
\bar{\lambda}_c^{nnt}(x) =&\Lambda_0\left(-\infty,-\psi_c(x),-\infty,-\psi_h(x,1);\rho(x)\right),\\
\bar{\lambda}_h^{cnt}(x) =&\Lambda_0\left(-\psi_c(x),\infty,\tfrac{\psi_h(x,1)-\psi_c(x)}{\sqrt{2(1-\rho(x))}},\infty;\sqrt{\tfrac{1-\rho(x)}{2}}\right)\\
&-\sqrt{2(1-\rho(x))}\Lambda_0\left(\tfrac{\psi_h(x,1)-\psi_c(x)}{\sqrt{2(1-\rho(x))}},\infty,-\psi_c(x),\infty;\sqrt{\tfrac{1-\rho(x)}{2}}\right),\\
\bar{\lambda}_c^{cnt}(x) =&\Lambda_0\left(-\psi_c(x),\infty,\tfrac{\psi_h(x,1)-\psi_c(x)}{\sqrt{2(1-\rho(x))}},\infty;\sqrt{\tfrac{1-\rho(x)}{2}}\right),\\
\bar{\lambda}_h^{at}(x) =&\Lambda_0\left(-\psi_h(x,0),\infty,\tfrac{\psi_c(x)-\psi_h(x,0)}{\sqrt{2(1-\rho(x))}},\infty;\sqrt{\tfrac{1-\rho(x)}{2}}\right),\\
\bar{\lambda}_c^{at}(x) =&\Lambda_0\left(-\psi_h(x,0),\infty,\tfrac{\psi_c(x)-\psi_h(x,0)}{\sqrt{2(1-\rho(x))}},\infty;\sqrt{\tfrac{1-\rho(x)}{2}}\right)\\
&-\sqrt{2(1-\rho(x))}\Lambda_0\left(\tfrac{\psi_c(x)-\psi_h(x,0)}{\sqrt{2(1-\rho(x))}},\infty,-\psi_h(x,0),\infty;\sqrt{\tfrac{1-\rho(x)}{2}}\right).
\end{aligned}
$$

The $\bar{\lambda}_d^g(x)$ are functions of the multinomial probit parameters. By the arguments above, this means that they can also be expressed as implicit functions of the choice probabilities, $\pi_d(x,z)$.

### Identification of selection coefficients

First consider identification of the selection coefficients for outcome $n$, $\gamma_n^h$ and $\gamma_n^c$. We can write the two identified means for the $n$ equation as

$$\mu_n^{nc}(x) = \mu_n(x) + \gamma_n^h\bar{\lambda}_h^{nc}(x) + \gamma_n^c\bar{\lambda}_c^{nc}(x),$$

$$\mu_n^{nnt}(x) = \mu_n(x) + \gamma_n^h\bar{\lambda}_h^{nnt}(x) + \gamma_n^c\bar{\lambda}_c^{nnt}(x)\ .$$

Differencing these two equations yields

$$\mu_n^{nc}(x) - \mu_n^{nnt}(x) = \gamma_n^h\left(\bar{\lambda}_h^{nc}(x) - \bar{\lambda}_h^{nnt}(x)\right) + \gamma_n^c\left(\bar{\lambda}_c^{nc}(x) - \bar{\lambda}_c^{nnt}(x)\right). \tag{26}$$

Evaluating equation (26) at $x=0$ and $x=1$ and then solving for $\gamma_n^h$ and $\gamma_n^c$ yields

$$\gamma_n^h = \frac{\left(\bar{\lambda}_c^{nc}(1) - \bar{\lambda}_c^{nnt}(1)\right)\left(\mu_n^{nc}(0) - \mu_n^{nnt}(0)\right) - \left(\bar{\lambda}_c^{nc}(0) - \bar{\lambda}_c^{nnt}(0)\right)\left(\mu_n^{nc}(1) - \mu_n^{nnt}(1)\right)}{\left(\bar{\lambda}_c^{nc}(1) - \bar{\lambda}_c^{nnt}(1)\right)\left(\bar{\lambda}_h^{nc}(0) - \bar{\lambda}_h^{nnt}(0)\right) - \left(\bar{\lambda}_c^{nc}(0) - \bar{\lambda}_c^{nnt}(0)\right)\left(\bar{\lambda}_h^{nc}(1) - \bar{\lambda}_h^{nnt}(1)\right)},$$

$$\gamma_n^c = \frac{\left(\bar{\lambda}_h^{nc}(1) - \bar{\lambda}_h^{nnt}(1)\right)\left(\mu_n^{nc}(0) - \mu_n^{nnt}(0)\right) - \left(\bar{\lambda}_h^{nc}(0) - \bar{\lambda}_h^{nnt}(0)\right)\left(\mu_n^{nc}(1) - \mu_n^{nnt}(1)\right)}{\left(\bar{\lambda}_c^{nc}(1) - \bar{\lambda}_c^{nnt}(1)\right)\left(\bar{\lambda}_h^{nc}(0) - \bar{\lambda}_h^{nnt}(0)\right) - \left(\bar{\lambda}_c^{nc}(0) - \bar{\lambda}_c^{nnt}(0)\right)\left(\bar{\lambda}_h^{nc}(1) - \bar{\lambda}_h^{nnt}(1)\right)}.$$

$\gamma_n^h$ and $\gamma_n^c$ are identified as long as the denominators of these ratios are not zero. This requires $X_i$ to shift utilities for $h$ and $c$, so that $(\bar{\lambda}_d^{nc}(x) - \bar{\lambda}_d^{nnt}(x))$ varies with $x$.

A similar argument shows identification of the selection coefficients for the $c$ equation. By evaluating $\mu_c^{cc}(x) - \mu_c^{cnt}(x)$ at $x = 0$ and $x = 1$, we obtain

$$\gamma_c^h = \frac{\left(\bar{\lambda}_c^{cc}(1) - \bar{\lambda}_c^{cnt}(1)\right)\left(\mu_c^{cc}(0) - \mu_c^{cnt}(0)\right) - \left(\bar{\lambda}_c^{cc}(0) - \bar{\lambda}_c^{cnt}(0)\right)\left(\mu_c^{cc}(1) - \mu_c^{cnt}(1)\right)}{\left(\bar{\lambda}_c^{cc}(1) - \bar{\lambda}_c^{cnt}(1)\right)\left(\bar{\lambda}_h^{cc}(0) - \bar{\lambda}_h^{cnt}(0)\right) - \left(\bar{\lambda}_c^{cc}(0) - \bar{\lambda}_c^{cnt}(0)\right)\left(\bar{\lambda}_h^{cc}(1) - \bar{\lambda}_h^{cnt}(1)\right)},$$

$$\gamma_c^c = \frac{\left(\bar{\lambda}_h^{cc}(1) - \bar{\lambda}_h^{cnt}(1)\right)\left(\mu_c^{cc}(0) - \mu_c^{cnt}(0)\right) - \left(\bar{\lambda}_h^{cc}(0) - \bar{\lambda}_h^{cnt}(0)\right)\left(\mu_c^{cc}(1) - \mu_c^{cnt}(1)\right)}{\left(\bar{\lambda}_c^{cc}(1) - \bar{\lambda}_c^{cnt}(1)\right)\left(\bar{\lambda}_h^{cc}(0) - \bar{\lambda}_h^{cnt}(0)\right) - \left(\bar{\lambda}_c^{cc}(0) - \bar{\lambda}_c^{cnt}(0)\right)\left(\bar{\lambda}_h^{cc}(1) - \bar{\lambda}_h^{cnt}(1)\right)}.$$

Identification of the coefficients in the $h$ equation is achieved by differencing the treated complier mean and the always taker mean, $\mu_h^c(x) - \mu_h^{at}(x)$, and evaluating this difference at zero and one. This gives

$$\gamma_h^h = \frac{\left(\bar{\lambda}_c^{c}(1) - \bar{\lambda}_c^{at}(1)\right)\left(\mu_h^{c}(0) - \mu_h^{at}(0)\right) - \left(\bar{\lambda}_c^{c}(0) - \bar{\lambda}_c^{at}(0)\right)\left(\mu_h^{c}(1) - \mu_h^{at}(1)\right)}{\left(\bar{\lambda}_c^{c}(1) - \bar{\lambda}_c^{at}(1)\right)\left(\bar{\lambda}_h^{c}(0) - \bar{\lambda}_h^{at}(0)\right) - \left(\bar{\lambda}_c^{c}(0) - \bar{\lambda}_c^{at}(0)\right)\left(\bar{\lambda}_h^{c}(1) - \bar{\lambda}_h^{at}(1)\right)},$$

$$\gamma_h^c = \frac{\left(\bar{\lambda}_h^{c}(1) - \bar{\lambda}_h^{at}(1)\right)\left(\mu_h^{c}(0) - \mu_h^{at}(0)\right) - \left(\bar{\lambda}_h^{c}(0) - \bar{\lambda}_h^{at}(0)\right)\left(\mu_h^{c}(1) - \mu_h^{at}(1)\right)}{\left(\bar{\lambda}_c^{c}(1) - \bar{\lambda}_c^{at}(1)\right)\left(\bar{\lambda}_h^{c}(0) - \bar{\lambda}_h^{at}(0)\right) - \left(\bar{\lambda}_c^{c}(0) - \bar{\lambda}_c^{at}(0)\right)\left(\bar{\lambda}_h^{c}(1) - \bar{\lambda}_h^{at}(1)\right)},$$

where

$$\bar{\lambda}_d^c(x) = S_c(x) \cdot \bar{\lambda}_d^{cc}(x) + (1 - S_c(x)) \cdot \bar{\lambda}_d^{nc}(x), \ d \in \{h, c\},$$

and

$$S_c(x) = -\frac{\pi_c(x,1) - \pi_c(x,0)}{\pi_h(x,1) - \pi_h(x,0)}.$$

This expression reflects the fact that the treated complier mean is a weighted average of means for $n$- and $c$-compliers.

## Identification of Average Treatment Effects

After identifying the selection parameters, it is straightforward to recover the population average potential outcomes $\mu_d(x)$:

$$\mu_h(x) = \mu_h^{at}(x) - \gamma_h^h \bar{\lambda}_h^{at}(x) - \gamma_h^c \bar{\lambda}_c^{at}(x),$$

$$\mu_c(x) = \mu_c^{cnt}(x) - \gamma_c^h \bar{\lambda}_h^{cnt}(x) - \gamma_c^c \bar{\lambda}_c^{cnt}(x),$$

$$\mu_n(x) = \mu_n^{nnt}(x) - \gamma_n^h \bar{\lambda}_h^{nnt}(x) - \gamma_n^c \bar{\lambda}_c^{nnt}(x).$$

These expressions adjust the always-taker, $c$-never taker, and $n$-never taker means for non-random selection.

### Identification of subLATEs

We can then construct $x$-specific subLATEs as follows:

$$LATE_{nh}(x) = \mu_h(x) - \mu_n(x) + \left(\gamma_h^h - \gamma_n^h\right) \bar{\lambda}_h^{nc}(x) + \left(\gamma_h^c - \gamma_n^c\right) \bar{\lambda}_c^{nc}(x),$$

$$LATE_{ch}(x) = \mu_h(x) - \mu_c(x) + \left(\gamma_h^h - \gamma_c^h\right) \bar{\lambda}_h^{cc}(x) + \left(\gamma_h^c - \gamma_c^c\right) \bar{\lambda}_c^{cc}(x).$$

Recall that the the selection coefficients and mean potential outcomes depend on a set of $\mu_d^g(x)$ that are nonparametrically identified, along with the $\bar{\lambda}_d^g(x)$, which depend in turn on conditional choice probabilities. This argument shows that under the semiparametric additive separability restriction (25), we can identify the selection model and the subLATEs by combining these moments.

Figure 1: Compliance Patterns in the HSIS

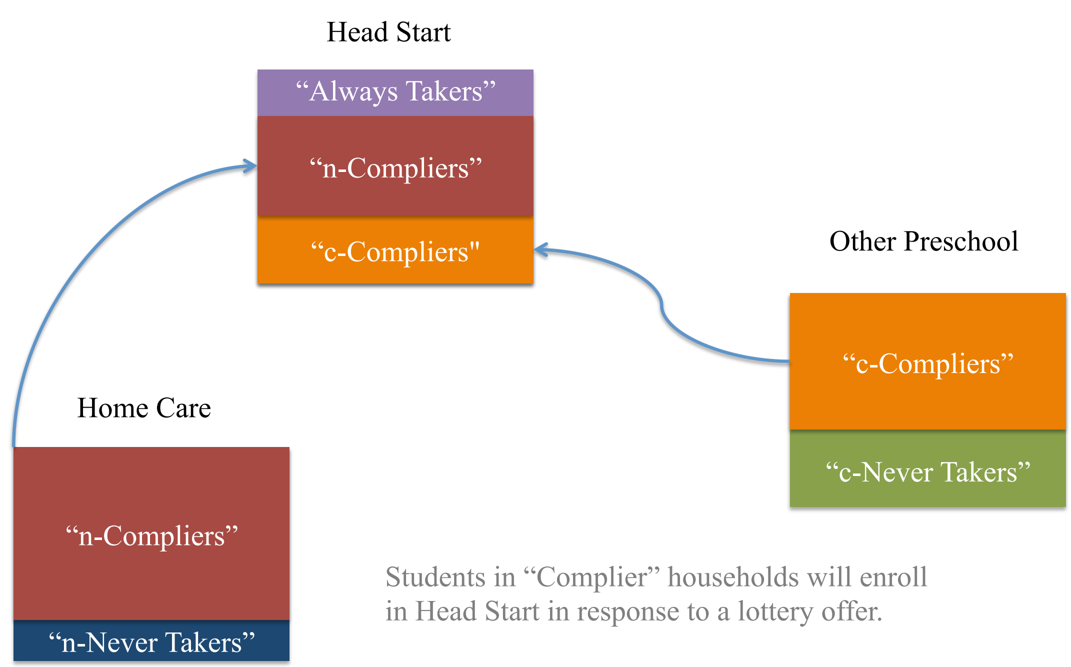
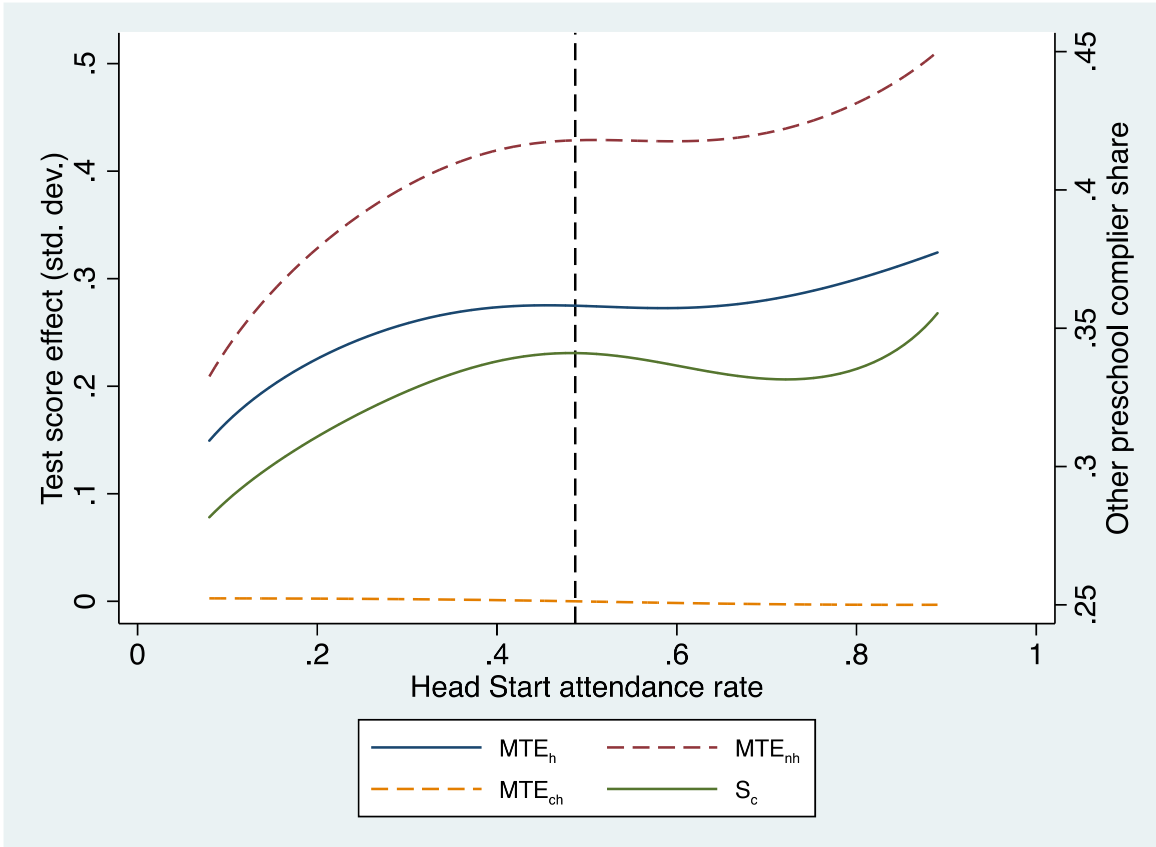
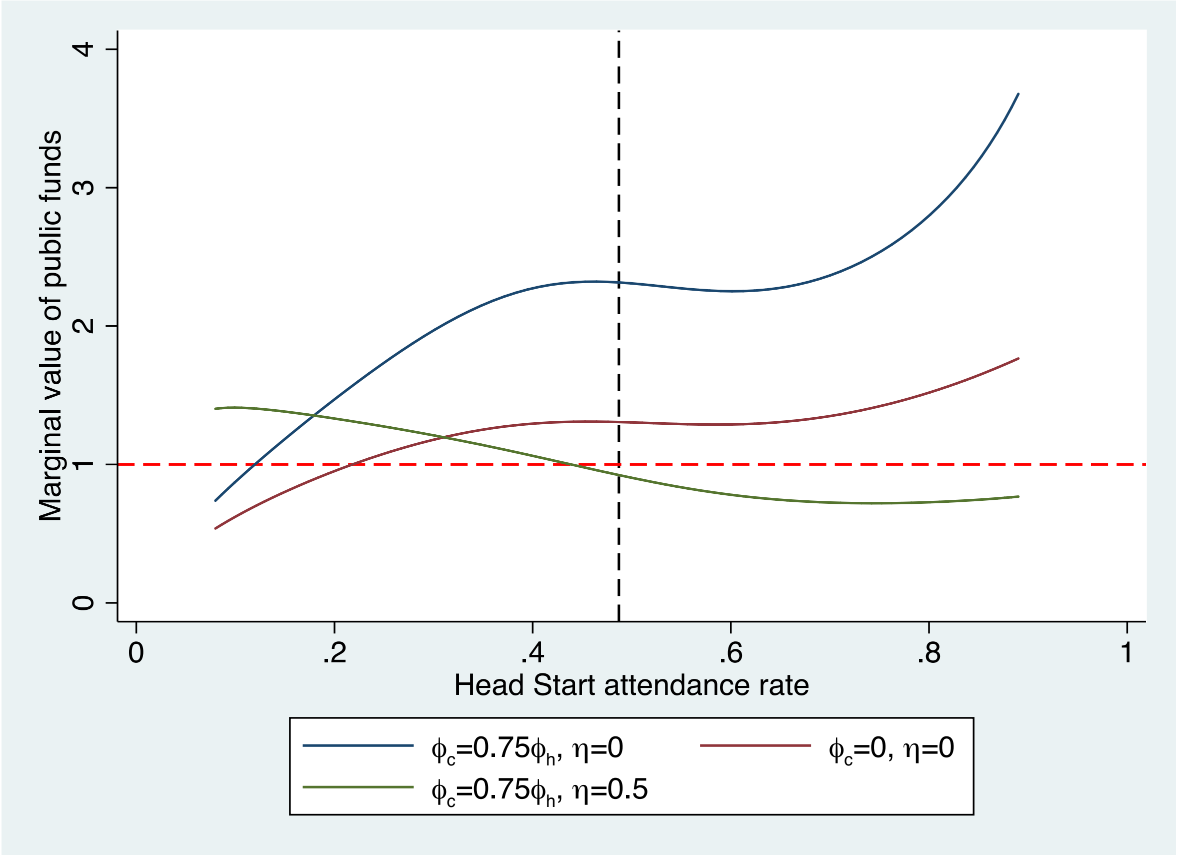


Figure 2: Test Score Effects of Structural Reforms




Notes: This figure plots predicted test score effects and complier shares for various values of the program feature *f*, which shifts the utility of Head Start attendance. The horizontal axis shows the Head Start attendance rate at each *f*. The left axis measures test score effects. $MTE_h$ (blue) is the average effect for marginal students, while $MTE_{nh}$ and $MTE_{ch}$ (red and yellow) are effects for subgroups of marginal students drawn from home care and other preschools. The right axis measures $S_c$ (greene), the share of marginal students drawn from other preschools.

Figure 3: Marginal Value of Public Funds for Structural Reforms




Notes: This figure plots the predicted marginal value of public funds for structural reforms of Head Start. These predictions use the same parameter calibrations as Table 6. The red curve sets the cost of competing programs and the elasticity of Head Start's cost with respect to program scale equal to zero. The blue curve sets the cost of competing programs equal to three-fourths of Head Start's cost. The green curve sets the cost elasticity to 0.5.

Table 1: Descriptive Statistics

| | By offer status | | By preschool choice | | |
|---|---|---|---|---|---|
| | Non-offered mean | Offer differential | Head Start | Other centers | No preschool |
| Variable | (1) | (2) | (3) | (4) | (5) |
| Male | 0.505 | -0.011 | 0.501 | 0.506 | 0.492 |
| | | (0.019) | | | |
| Black | 0.298 | 0.010 | 0.317 | 0.353 | 0.250 |
| | | (0.010) | | | |
| Hispanic | 0.369 | 0.007 | 0.380 | 0.354 | 0.373 |
| | | (0.010) | | | |
| Teen mother | 0.174 | -0.015 | 0.159 | 0.169 | 0.176 |
| | | (0.014) | | | |
| Mother married | 0.448 | -0.011 | 0.439 | 0.420 | 0.460 |
| | | (0.017) | | | |
| Both parents in household | 0.488 | 0.009 | 0.497 | 0.468 | 0.499 |
| | | (0.017) | | | |
| Mother is high school dropout | 0.397 | -0.029 | 0.377 | 0.322 | 0.426 |
| | | (0.017) | | | |
| Mother attended some college | 0.281 | 0.017 | 0.293 | 0.342 | 0.253 |
| | | (0.016) | | | |
| Test language is not English | 0.239 | 0.016 | 0.268 | 0.223 | 0.231 |
| | | (0.011) | | | |
| Home language is not English | 0.273 | 0.014 | 0.296 | 0.274 | 0.260 |
| | | (0.011) | | | |
| Special education | 0.108 | 0.028 | 0.134 | 0.145 | 0.091 |
| | | (0.011) | | | |
| Only child | 0.139 | 0.022 | 0.151 | 0.190 | 0.123 |
| | | (0.012) | | | |
| Income (fraction of FPL)* | 0.896 | 0.000 | 0.892 | 0.983 | 0.851 |
| | | (0.024) | | | |
| Age 4 cohort | 0.451 | -0.003 | 0.426 | 0.567 | 0.413 |
| | | (0.012) | | | |
| Baseline summary index | 0.012 | -0.009 | -0.001 | 0.106 | -0.040 |
| | | (0.027) | | | |
| Center provides transportation | 0.604 | 0.002 | 0.586 | 0.614 | 0.628 |
| | | (0.005) | | | |
| Center quality index | 0.678 | -0.001 | 0.679 | 0.681 | 0.673 |
| | | (0.003) | | | |
| Joint *p*-value | | 0.268 | | | |
| N | 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 joint *p*-value is from a test of the hypothesis that all coefficients equal zero.
*Household income is missing for 19 percent of observations. Missing values are excluded in statistics for income.

Table 2: 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 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Year 1 | 0.194 | 0.699 | 0.278 | 0.141 | 0.664 | 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 | 0.089 | 0.356 | 0.249 | -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) | (0.048) |
| N | | 1760 | | | 1416 | | | 3176 | |
| Year 3 | -0.008 | 0.366 | -0.023 | 0.055 | 0.666 | 0.082 | 0.020 | 0.500 | 0.040 |
| | (0.031) | (0.028) | (0.084) | (0.040) | (0.025) | (0.060) | (0.025) | (0.020) | (0.050) |
| N | | 1659 | | | 1336 | | | 2995 | |
| Year 4 | 0.039 | 0.344 | 0.114 | | - | | | - | |
| | (0.033) | (0.028) | (0.097) | | | | | | |
| 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, teen mother, mother marital status, presence of both parents in the home, family size, special education status, test language, home language, income quartile dummies, 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 3: Preschool Choices by Year, Cohort, and Offer Status

| | | Offered | | | Not offered | | | Other center complier share |
|---|---|---|---|---|---|---|---|---|
| | | Head Start | Other centers | No preschool | Head Start | Other centers | No preschool | |
| 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 4: Funding Sources

| Largest funding source | Head Start (1) | Other centers (2) | Other centers attended by $c$ -> $h$ compliers (3) |
|---|---|---|---|
| Head Start | 0.842 | 0.027 | 0.038 |
| Parent fees | 0.004 | 0.153 | 0.191 |
| Child and adult care food program | 0.011 | 0.026 | 0.019 |
| State pre-K program | 0.004 | 0.182 | 0.155 |
| Child care subsidies | 0.013 | 0.097 | 0.107 |
| Other funding or support | 0.022 | 0.118 | 0.113 |
| No funding or support | 0.000 | 0.003 | 0.001 |
| Missing | 0.105 | 0.394 | 0.375 |

Notes: This table reports largest funding sources for Head Start and other preschool centers. Reported funding sources come from interviews with childcare center directors. Column (3) reports funding sources for other preschool centers attended by non-offered compliers who would be induced to attend Head Start by an experimental offer. Estimates in this column are produced using the methods in Appendix D.

Table 5: Characteristics of Head Start and Competing Preschool Centers

| | Head Start<br>(1) | Other centers<br>(2) | Other centers attended by *c* -> *h* compliers<br>(3) |
|---|---|---|---|
| Transportation provided | 0.629 | 0.383 | 0.324 |
| Quality index | 0.702 | 0.453 | 0.446 |
| Fraction of staff with bachelor's degree | 0.345 | 0.527 | 0.491 |
| Fraction of staff with teaching license | 0.113 | 0.260 | 0.247 |
| Center director experience | 18.2 | 12.2 | 12.6 |
| Student/staff ratio | 6.80 | 8.24 | 8.54 |
| Full day service | 0.637 | 0.735 | 0.698 |
| More than three home visits per year | 0.192 | 0.073 | 0.072 |
| N | 1848 | 366 | |

Notes: This table reports center characteristics obtained from a survey of center directors. Column (1) shows characteristics of Head Start centers attended by children in the HSIS sample, while column (2) shows characteristics of other preschool centers. Column (3) 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 this column are produced using the methods in Appendix D.

Table 6: Benefits and Costs of Head Start

| Parameter (1) | Description (2) | Value (3) | Source (4) |
|---|---|---|---|
| | *Panel A. Parameter values* | | |
| $p$ | Effect of a 1 SD increase in test scores on earnings | $0.1\bar{e}$ | Table A3 |
| $e_{US}$ | 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 |
| $\tau$ | 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 |
| $\phi_h$ | Marginal cost of enrollment in Head Start | $8,000 | Head Start program facts |
| $\phi_c$ | Marginal cost of enrollment in other preschools | $0 | Naïve assumption: $\phi_c = 0$ |
| | | $4,000 | Pessimistic assumption: $\phi_c = 0.5\phi_h$ |
| | | $7,500 | 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 |
| | | $2,991 | Preferred assumption |
| *MVPF* | **Marginal value of public funds** | 1.10 (0.22)<br>$p$-value = 0.1<br>Breakeven $p/\bar{e} = 0.09$ (0.01) | *NMB*/*MFC* (s.e.), naïve assumption |
| | | 1.50 (0.34)<br>$p$-value = 0.00<br>Breakeven $p/\bar{e} = 0.08$ (0.01) | Pessimistic assumption |
| | | 1.84 (0.47)<br>$p$-value = 0.00<br>Breakeven $p/\bar{e} = 0.07$ (0.01) | Preferred assumption |

Notes: This table reports results of cost/benefit calculations for Head Start. Estimated 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 7: Mean Potential Outcomes for Subpopulations

| | Probability<br>(1) | $E[Y(h)]$<br>(3) | $E[Y(c)]$<br>(5) | $E[Y(n)]$<br>(7) |
|---|---|---|---|---|
| $n$-compliers | 0.454 | - | - | -0.078 |
| $c$-compliers | 0.232 | - | 0.107 | - |
| All compliers | 0.686 | 0.233 | - | - |
| $n$-never takers | 0.095 | - | - | -0.035 |
| $c$-never takers | 0.083 | - | 0.316 | - |
| Always takers | 0.136 | -0.028 | - | - |

Notes: This table reports estimates of mean potential outcomes for subpopulations defined by response to the HSIS experimental offer. Mean potential outcomes for compliers are estimated via the methods described in Appendix D.

Table 8: Multinomial Probit Estimates

| | Head Start utility | | | |
|---|---|---|---|---|
| | Main effect | Offer interaction | Other center utility | Arctanh $\rho$ |
| | (1) | (2) | (3) | (4) |
| Constant | -0.910 | 2.127 | -0.375 | 0.303 |
| | (0.075) | (0.087) | (0.054) | (0.067) |
| Transportation | -0.536 | 0.708 | -0.042 | -0.172 |
| | (0.168) | (0.194) | (0.123) | (0.160) |
| Above-median quality | -0.343 | 0.548 | 0.037 | 0.010 |
| | (0.157) | (0.181) | (0.107) | (0.150) |
| Mother's education | -0.035 | 0.145 | 0.121 | -0.166 |
| | (0.075) | (0.089) | (0.060) | (0.082) |
| Income above FPL | 0.270 | -0.337 | 0.149 | 0.050 |
| | (0.152) | (0.157) | (0.140) | (0.173) |
| Age 4 | 0.068 | -0.143 | 0.469 | 0.103 |
| | (0.128) | (0.148) | (0.106) | (0.147) |
| *P*-value | 0.000 | 0.000 | 0.000 | 0.411 |
| Log-likelihood | | -2587.3 | | |

Notes: This table reports simulated maximum likelihood estimates of a multinomial probit model of preschool choice. The likelihood is evaluated using the GHK simulator, and likelihood contributions are weighted by the reciprocal of the probability of experimental assignments. *P*-values are from tests that all coefficients in a column except the constant term are zero. The Head Start and other center utilities also include the main effects of gender, race, assignment cohort, teen mother, mother's education, mother's marital status, presence of both parents, an only child dummy, special education, test language, home language, dummies for quartiles of family income and missing income, an indicator for whether the Head Start center provides transportation, the Head Start quality index, and a third-order polynomial in baseline test scores. Standard errors are clustered at the center level.

Table 9: Selection-corrected Estimates of Preschool Effects

| | | Least squares | | Two-step | | | |
|---|---|---|---|---|---|---|---|
| | | No controls | Baseline controls | Unrestricted | Covs. restricted | Selection restricted | ATE restricted |
| Parameter | Description | (1) | (2) | (3) | (4) | (5) | (6) |
| $\theta_h^0$ - $\theta_n^0$ | Effect of Head Start relative to no preschool | 0.202<br>(0.037) | 0.214<br>(0.022) | 0.722<br>(0.198) | 0.437<br>(0.120) | 0.456<br>(0.116) | 0.473<br>(0.110) |
| $\theta_c^0$ - $\theta_n^0$ | Effect of other preschools relative to no preschool | 0.262<br>(0.052) | 0.149<br>(0.033) | 0.406<br>(0.596) | 0.172<br>(0.274) | 0.372<br>(0.234) | 0.473<br>(0.110) |
| $\gamma_h^h$ | Coefficient on Head Start taste in Head Start outcome equation | - | - | -0.147<br>(0.051) | -0.150<br>(0.051) | -0.132<br>(0.046) | -0.137<br>(0.044) |
| $\gamma_c^h$ | Coefficient on Head Start taste in other preschool outcome equation | | - | -0.008<br>(0.182) | -0.029<br>(0.082) | -0.132<br>(0.046) | -0.137<br>(0.044) |
| $\gamma_n^h$ | Coefficient on Head Start taste in no preschool outcome equation | - | - | 0.084<br>(0.077) | -0.004<br>(0.056) | 0.003<br>(0.055) | 0.008<br>(0.054) |
| $\gamma_h^c$ | Coefficient on other preschool taste in Head Start outcome equation | - | - | 0.030<br>(0.341) | -0.073<br>(0.329) | -0.037<br>(0.151) | -0.109<br>(0.032) |
| $\gamma_c^c$ | Coefficient on other preschool taste in other preschool outcome equation | - | - | 0.163<br>(0.509) | 0.123<br>(0.186) | -0.037<br>(0.151) | -0.109<br>(0.032) |
| $\gamma_n^c$ | Coefficient on other preschool taste in no preschool outcome equation | - | - | -0.754<br>(0.349) | -0.242<br>(0.220) | -0.286<br>(0.208) | -0.319<br>(0.196) |
| | *P*-value for all restrictions | - | - | - | 0.762 | 0.661 | 0.539 |
| | *P*-value for additional restrictions | - | - | - | 0.762 | 0.788 | 0.687 |
| | *P*-value: No selection on gains | - | - | 0.128 | 0.204 | 0.129 | 0.093 |
| | *P*-value: No selection on gains or levels | - | - | 0.014 | 0.029 | 0.009 | 0.001 |

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 and controls, separately for children attending Head Start, other preschools, and no preschool. The first two rows report differences in intercepts between Head Start and no preschool, and other preschools and no preschool. Column (1) shows estimates with no controls. Column (2) adds controls for the same baseline covariates used in Table 8. Covariates are de-meaned in the estimation sample, so that differences in intercepts can be interpreted as effects at the mean. Column (3) adds selection-correction terms. Column (4) restricts coefficients on the covariates to be the same in each care alternative, except transportation, above-median quality, mother's education, income above the poverty line, age 4, baseline score, and race.  Column (5) restricts the coefficient on the Head Start utility to be the same in the Head Start and other center equations, and similarly for the other center utility. Column (6) restricts the intercepts in the Head Start and other center equations to be the same. Standard errors are boostrapped and clustered at the center level.

Table 10: 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 | 0.447 | - | 0.307 | - | 0.306 | -0.078 | -0.068 |
| $c$-compliers | 0.232 | 0.237 | - | 0.151 | 0.107 | 0.153 | - | -0.497 |
| All compliers | 0.686 | 0.683 | 0.233 | 0.253 | - | 0.253 | - | -0.217 |
| $n$-never takers | 0.095 | 0.097 | - | 0.500 | - | 0.500 | -0.035 | -0.057 |
| $c$-never takers | 0.083 | 0.081 | - | 0.316 | 0.316 | 0.320 | - | -0.541 |
| Always takers | 0.136 | 0.138 | -0.028 | -0.042 | - | -0.045 | - | -0.348 |
| Full population | 1 | 1 | - | 0.228 | - | 0.228 | - | -0.245 |
| $p$-value: IV = Two-step | 0.354 | | 0.217 | | 0.033 | | 0.869 | |
| $p$-value for all moments | | | | 0.126 | | | | |

Notes: This table compares nonparametric estimates of mean potential outcomes for subpopulations to estimates implied by the two-step model in column (6) of Table 9.

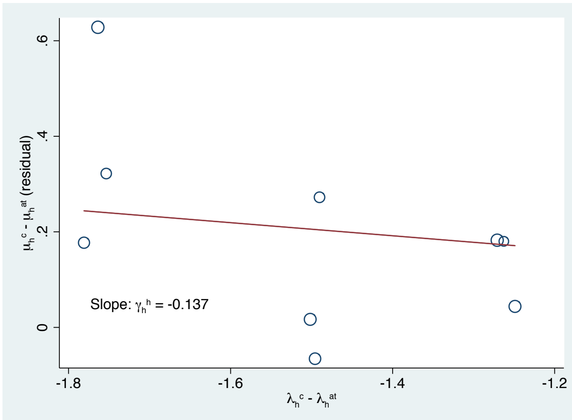
Table 11: Treatment Effects for Subpopulations

| | | Two-step | | | |
|---|---|---|---|---|---|
| | IV | Unrestricted | Covariates restricted | Selection restricted | ATE restricted |
| Parameter | (1) | (2) | (3) | (4) | (5) |
| LATE | 0.247 | 0.260 | 0.256 | 0.247 | 0.245 |
| | (0.031) | (0.036) | (0.032) | (0.031) | (0.030) |
| *n* -> *h* subLATE | - | 0.315 | 0.363 | 0.336 | 0.375 |
| | | (0.179) | (0.174) | (0.093) | (0.047) |
| *c* -> *h* subLATE | - | 0.156 | 0.053 | 0.079 | -0.002 |
| | | (0.323) | (0.310) | (0.161) | (0.013) |
| *n* -> *h* ATE | - | 0.722 | 0.437 | 0.456 | 0.473 |
| | | (0.198) | (0.120) | (0.116) | (0.110) |
| *c* -> *h* ATE | - | 0.316 | 0.265 | 0.084 | 0 |
| | | (0.577) | (0.209) | (0.168) | - |

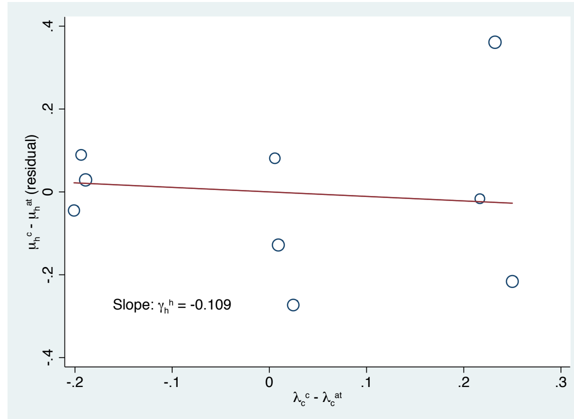
Notes: This table reports estimates of treatment effects for subpopulations. Column (1) reports an IV estimate of the effect of Head Start. Columns (2)-(5) show estimates of treatment effects computed from two-step models. Standard errors are boostrapped and clustered at the center level.

Figure A1: Identification of the Selection Model

*Panel A. Equation for Y(h)*

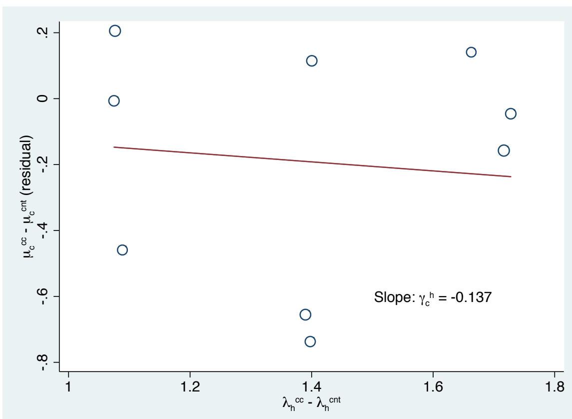



Difference in $Y(h)$ against difference in $v_h$, compliers vs. always takers

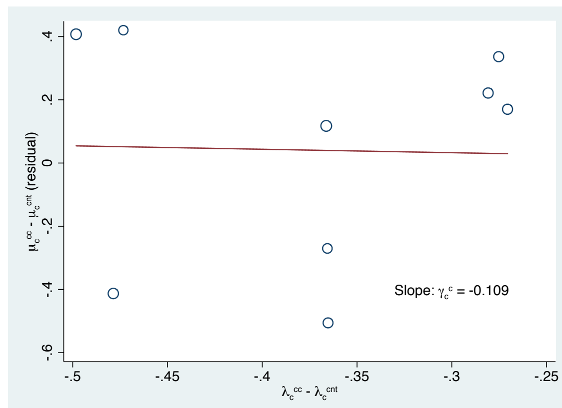



Difference in $Y(h)$ against difference in $v_c$, compliers vs. always takers

*Panel B. Equation for Y(c)*

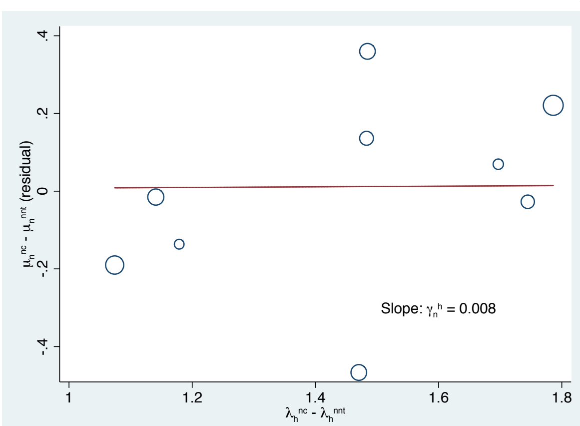



Difference in $Y(c)$ against difference in $v_h$, c-compliers vs. $c$-never takers

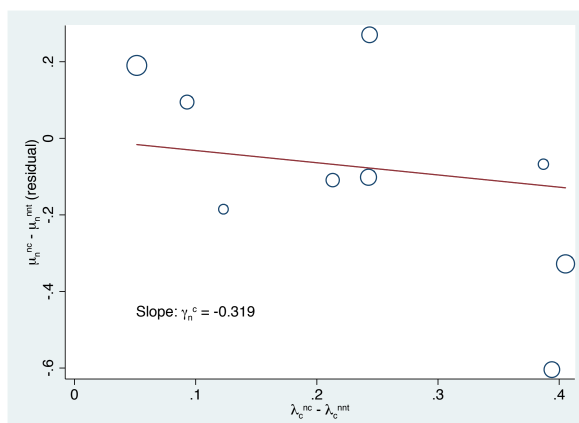



Difference in $Y(c)$ against difference in $v_c$, $c$-compliers vs. $c$-never takers

*Panel C. Equation for Y(n)*




Difference in $Y(n)$ against difference in $v_h$, n-compliers vs. $n$-never takers




Difference in $Y(n)$ against difference in $v_c$, $n$-compliers vs. $n$-never takers

Notes: This figure illustrates identification of the selection model by comparing differences in mean potential outcomes and differences in average unobserved tastes across compliance groups. Panel A shows differences between compliers and always takers in Head Start. Panel B shows differences between $c$-compliers and $c$-never takers in other preschools. Panel C shows differences between $n$-compliers and $n$-never takers in home care. In each panel, average differences in tastes are computed conditional on each observation's covariates, and the sample is split into nine groups constructed by interacting terciles of each taste. Explicit formulas for differences in tastes are provided in the Appendix. Scatter plots show average residual differences in potential outcomes within each cell against average differences in tastes. Residuals are constructed by subtracting off coefficients from the restricted model in Table 9. Lines are plotted through the origin with slopes equal to the relevant coefficient from Table 9.

Table A1: 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 A2: 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 A3: Estimates of Test Score and Earnings Impacts

| Study | Intervention (1) | Test score effect (2) | Earnings effect (3) | Ratio (4) |
|---|---|---|---|---|
| Chetty et al. (2011) | Tennessee STAR (1 s.d. of class quality, kindergarten)[a] | 0.024 | 0.003 | 0.131 |
| | OLS relationship (w/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 relationship (w/controls, grades 3-8)[d] | 1.0 | 0.12 | 0.12 |
| Currie and Thomas (1995), Garces et al. (2002) | Head Start (whites, mother fixed effects)[e] | 0.217 | 0.566 | 2.61 |
| | Head Start (blacks, mother fixed effects)[f] | 0.009 | 0.073 | 8.11 |
| Heckman et al. (2010) | Perry Preschool Project (males)[g] | 0.787 | 0.189 | 0.240 |
| | Perry Preschool Project (females)[h] | 0.980 | 0.286 | 0.292 |
| Murnane et al. (1995) | OLS relationship (males, w/controls, grade 12)[i] | 1.0 | 0.077 | 0.077 |
| | OLS relationship (females, w/controls, grade 12)[j] | 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]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.

[f]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.

[g]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).

[h]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).

[i]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).

[j]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 A4: Interacted Two-stage Least Squares Estimates

| | Single endogenous variable: | Two endogenous variables: | |
|---|---|---|---|
| | Head Start | Head Start | Other centers |
| Model | (1) | (2) | (3) |
| Just-identified | 0.247 | - | - |
| | (0.031) | | |
| Overidentified | 0.238 | 0.360 | 0.358 |
| | (0.030) | (0.148) | (0.419) |
| First-stage $F$ | 421.7 | 19.4 | 1.9 |
| Overid. $p$-value | 0.048 | 0.038 | |

Notes: This table reports two-stage least squares estimates of the effects of Head Start and other preschool centers in Spring 2003. Column (1) and shows estimates treating Head Start as the endogenous variable. Columns (2) and (3) show estimates of a model treating Head Start and other preschools as separate endogenous variables. The just-identified model instruments with the Head Start offer. Overidentified models instrument with the offer interacted with transportation, above-median center quality, above-median income, age 4, and mother's education. All models weight by the reciprocal of the probability of a child's experimental assignment, and control for baseline covariates. Standard errors are clustered at the center level. $F$-statistics are Angrist/Pischke (2009) partial $F$'s.

Table A5: Benefits and Costs of Head Start when Competing Preschools are Rationed

| Parameter<br>(1) | Description<br>(2) | Value<br>(3) | Source<br>(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 |
| | | 0.375 | Homogeneity assumption: *n->c* subLATE equals *n->h* subLATE |
| | | 0.446 | Model-based prediction |
| *NMB* | Marginal benefit to Head Start population net of taxes | \$5,513 | $(1 - \tau)p(LATE_h+S_cLATE_{nc})$, naïve assumption |
| | | \$8,442 | Homogeneity assumption |
| | | \$8,997 | Model-based prediction |
| *MFC* | Marginal fiscal cost of Head Start enrollment | \$5,031 | $\phi_h - \tau p(LATE_h+S_cLATE_{nc})$, naïve assumption |
| | | \$3,454 | Homogeneity assumption |
| | | \$3,155 | Model-based prediction |
| *MVPF* | **Marginal value of public funds** | 1.10 | Naïve assumption |
| | | 2.44 | Homogeneity assumption |
| | | 2.85 | 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. Estimated parameter values are obtained from the sources listed in column (4).