IRLE
Institute for Research on Labor and Employment

IRLE WORKING PAPER
#123-14
December 2014

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

Patrick Kline and Christopher Walters

Cite as: Patrick Kline and Christopher Walters. (2014). "Evaluating Public Programs with Close Substitutes: The Case of Head Start". IRLE Working Paper No. 123-14. http://irle.berkeley.edu/workingpapers/123-14.pdf

Berkeley
UNIVERSITY OF CALIFORNIA

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

Patrick Kline
UC Berkeley/NBER

Christopher Walters*
UC Berkeley/NBER

December 2014

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 randomized Head Start Impact Study (HSIS), we show that Head Start draws a substantial share 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 social costs. A cost-benefit analysis demonstrates that accounting for the public savings associated with reduced enrollment in other subsidized preschools can reverse negative assessments of the program's social rate of return. Estimates from a semi-parametric selection model indicate that Head Start is about as effective at raising test scores as competing preschools and that its impacts are greater on children from families unlikely to participate in the program. Efforts to expand Head Start to new populations are therefore likely to boost the program's social 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, Magne Mogstad, and seminar participants at UC Berkeley, the University of Chicago, and Arizona State University for helpful comments. 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 the counterfactual for many participants involves receipt of similar services (Heckman et al., 2000). Likewise, neglecting program substitution patterns can lead to an overstatement of a program's net social costs if alternative programs are 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).[1]

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).

We critically reassess this conclusion in light of the fact that roughly one-third of the HSIS control group participated in alternate forms of preschool. Our study begins with a theoretical analysis that clarifies which parameters are (and are not) policy relevant when close program substitutes are present. We show that, for purposes of determining optimal program scale, the policy-relevant causal parameter is an average effect of participation relative to the next best alternative, regardless of whether that alternative is a competing program or nonparticipation. 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 experimental compliance, often thought to be a confounding limitation of practical experiments, is actually a virtue when the compliance patterns in the experiment replicate those found in the broader population.

Next, we show that a proper evaluation of a program's social costs must account for substitution patterns: if substitute programs receive public funds, an expansion of the target program generates cost savings in these programs, reducing net social costs. For example, in the polar case where all Head Start enrollees would have participated in other subsidized preschool programs with equivalent

[1] An additional 200,000 children participate in Early Head Start, which serves children under age 3.

social costs, the modest HSIS test score impacts would imply a "free lunch" can be obtained by shifting children out of competing programs and into Head Start. We conclude the theoretical analysis by considering structural reforms that alter features of the program other than its scale. 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 a randomized experiment.

Having established this theoretical groundwork, we use data from the HSIS to empirically assess the social returns to Head Start. The empirical analysis proceeds in three steps. First, we provide a nonparametric investigation of counterfactual preschool choices and Head Start's effects on test scores. We find that one third of compliers in the HSIS experiment would have participated in other forms of preschool had they not been lotteried into the program. Survey data on center administrators indicate that these compliers would have attended competing programs that draw heavily on public funding, which mitigates the net costs to government of enrolling them in Head Start. An analysis of test score effects replicates the fade-out pattern found in previous work and reveals 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 would suggest.

Second, we conduct a formal cost-benefit analysis of Head Start using results from Chetty et al. (2011) to convert short-run test score impacts into dollar equivalents. This analysis demonstrates that accounting for substitution from other socially costly programs is crucial for accurately assessing the social value of Head Start. We calculate social returns under several assumptions about the social costs of alternative preschools. Estimated social returns are negative when the costs of competing programs are set to zero but positive for more realistic values. Our preferred estimates suggest that Head Start pays for itself in increased earnings, with a projected social rate of return of approximately 15%. While this rate of return calculation relies on several difficult to verify assumptions, our analysis illustrates the quantitative point that ignoring program substitution can substantially distort an assessment of Head Start's cost-effectiveness.

Finally, we estimate a polychotomous selection model that allows us to separate the effects of Head Start and other preschools and to assess policy counterfactuals. We show that this model, which relies on some parametric restrictions, accurately reproduces patterns of treatment effect heterogeneity found in the experiment. Estimates of the model 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 the program experience larger test score gains. These results suggest that expanding the program to

new populations can boost its rate of return, provided that the proposed technology for increasing enrollment (e.g. improved transportation services) is not too costly.

The rest of the paper is structured as follows. Section 2 provides background on Head Start. Section 3 introduces a theoretical framework for assessing public programs with close substitutes. Section 4 describes the data. Section 5 reports nonparametric evidence on substitution patterns and test score effects. Section 6 provides a cost-benefit analysis of Head Start. Section 7 develops and estimates our econometric selection model. Section 8 simulates the effects of structural reforms. Section 9 discusses some important caveats to our analysis and concludes.

2 Background on Head Start

Head Start 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 program (TANF). 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, the distinction between Head Start and other public preschool programs may have more to do with the mix of funding sources being utilized than differences in education technology.

A large non-experimental literature suggests that Head Start produced large short- and long-run benefits for early cohorts of program participants. Currie and Thomas (1995) estimate the effects of Head Start by comparing program participants to their non-participant siblings. Their results show that Head Start participation boosted Peabody Picture and Vocabulary Test (PPVT) scores in elementary school for white children attending in the 1970s and 1980s, though not for black children. Garces et al. (2002) use the same methodology to show that Head Start increased educational attainment and earnings among whites and reduced crime among blacks. Similarly, Deming (2009) uses sibling comparisons to show that Head Start increased summary indices of

cognitive and non-cognitive skills for cohorts attending in the late 1980s. Deming concludes that Head Start produces approximately 80 percent of the benefits of the Perry Preschool project, a successful small-scale model program to which it is often compared, at 60 percent of the cost (Heckman et al., 2010a, 2010b, 2013). Ludwig and Miller (2007) use a regression discontinuity design based on a county-level cutoff in grant-writing aid to show that Head Start reduced child mortality in the early years of the program. Carneiro and Ginja (forthcoming) use a regression discontinuity design to evaluate more recent waves of the program and find long-run improvements in depression, obesity, and criminal activity among affected cohorts.

In contrast to these non-experimental estimates, the 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 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 popular media to motivate calls for dramatic restructuring or elimination of the program (Klein, 2011; Stossel, 2014).

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 program effects on parental engagement with children. Bitler et al. (2014) find that the experimental impacts are largest at bottom 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. Notably, Walters finds that effects are smaller for Head Start centers that draw more children from other preschools rather than home care, a finding we explore in more detail here.

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 comparisons. 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. Contemporaneous work by Feller et al. (2014) uses Bayesian principal stratification methods to estimate effects on subgroups of HSIS compliers drawn from other preschools and home care. Their results show positive effects for children drawn from home and negligible effects for children drawn from competing preschools. In Section 7, we provide further discussion of the methods and results in Feller et al. (2014) and their relationship to our approach. We turn now to developing a framework that allows us to formalize how the presence of program substitutes alters the evaluation problem.

3 A Model of Head Start Provision

In this section, we develop a simple model of Head Start participation with the goal of devising some efficiency criteria for provision of Head Start (or similar programs) when close program substitutes are available. Our model is highly stylized but serves to illustrate the point that partial knowledge of the technological parameters governing program outcomes is often sufficient to craft optimal policies provided that one can identify program substitution patterns.

There is a population of households, indexed by i, each with a single preschool-aged child. Households can participate in Head Start, a competing preschool program (e.g., state subsidized preschool), or care for their child at home. The government rations Head Start participation via program "offers" Z_i, which arrive at random 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, with h denoting enrollment in Head Start, c denoting enrollment in the competing program, and n denoting preschool non-participation. The argument $z \in \{0, 1\}$ indexes offer status: 0 denotes the absence of a Head Start offer while 1 denotes offer receipt. By assumption, 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. Household i's decision is given by:

$$D_i(z) = \underset{d \in \{h,c,n\}}{\arg\max} \; U_i(d, z). \tag{1}$$

Denote the probability of enrolling in option d given an offer by $\pi_d(1) = P(D_i(1) = d)$ and the probability of enrolling without an offer by $\pi_d(0) = P(D_i(0) = d)$. Enrollment in Head Start

is given by $N_h = \delta\pi_h(1) + (1-\delta)\pi_h(0)$. Likewise, enrollment in competing programs is $N_c = \delta\pi_c(1) + (1-\delta)\pi_c(0)$.

Every household has a set of potential test scores that would result under each of the three program alternatives, which we denote by the triple $\{Y_i(h), Y_i(c), Y_i(n)\}$. These potential outcomes are independent of the offer Z_i. Realized test scores can be written as a function of program participation decisions and whether a Head Start offer was received:

$$Y_i = \sum_{d\in\{h,c,n\}} Y_i(d)\left(1\{D_i(1)=d\}Z_i + 1\{D_i(0)=d\}(1-Z_i)\right).$$

Average realized test scores in the population are denoted $\bar{Y}$ and can be expressed as follows:

$$\begin{aligned}\bar{Y} &= E[Y_i|Z_i=1]\delta + E[Y_i|Z_i=0](1-\delta) \\ &= \sum_{d\in\{h,c,n\}} E[Y_i(d)|D_i(1)=d]\pi_d(1)\delta \\ &\quad + \sum_{d\in\{h,c,n\}} E[Y_i(d)|D_i(0)=d]\pi_d(0)(1-\delta),\end{aligned} \tag{2}$$

where the second line follows from the assumption that offers are assigned at random.

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). We assume that society values test score outcomes according to the money metric social criterion function:

$$W = g\left(\bar{Y}\right) - \phi_h N_h - \phi_c N_c, \tag{3}$$

where the function $g(.)$ is strictly increasing and maps average test scores $\bar{Y}$ into dollar equivalents. The scalar ϕ_h gives the social cost of providing Head Start services to an additional child. This social cost is measured in dollars and incorporates both the accounting cost to the government of funding the program and the deadweight costs associated with financing the program via distortionary taxes. Likewise, ϕ_c gives the social cost of providing competing preschool services to another student. If competing services are produced via a technology similar to Head Start and financed primarily via taxation, it is reasonable to expect $\phi_c \approx \phi_h$. We show empirically in Section 5 that a large fraction of competing preschools attended by Head Start-eligible children are financed by public subsidies.[2]

The social objective (3) reflects the paternalistic nature of the debate over early education

[2]The costs ϕ_h and ϕ_c are costs of Head Start and other preschools relative to preschool nonparticipation. Preschool participation may increase parents' labor supply, which could reduce social costs through increases in tax revenue or decreases in transfer payments. Appendix Table A1 shows that the experimental Head Start offer had no effect on the probability that a child's mother worked in Spring 2003. This suggests that labor supply responses are unlikely to importantly affect social costs in the Head Start-eligible population.

programs, which focuses almost entirely on the cost of boosting human capital for poor children with little regard to the impact on adults. By writing (3) 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. Moreover, we have abstracted from any positive externalities associated with schooling, such as reductions in crime (Lochner and Moretti, 2004; Heckman et al., 2010), that are not reflected in test scores. Incorporating such concerns into the social objective would be straightforward, but in the absence of widespread agreement over distributional motives and the size of any external effects, calibrating such a model would be difficult. Ignoring these issues yields a conservative lower bound estimate of the social return to Head Start that corresponds to the nature of public debate over the program.

Optimal program scale

We begin by considering the problem of choosing a maximum enrollment in Head Start, which is formally equivalent to choosing the rationing probability δ. From (2), the impact of a small change in the rationing probability on average test scores is:

$$\begin{aligned}\frac{d\bar{Y}}{d\delta} &= E\left[Y_i|Z_i=1\right]-E\left[Y_i|Z_i=0\right] \\ &= E\left[Y_i\left(h\right)-Y_i\left(D_i\left(0\right)\right)|D_i\left(1\right)=h,D_i\left(0\right)\neq h\right]\cdot P\left(D_i\left(1\right)=h,D_i\left(0\right)\neq h\right) \\ &\equiv LATE_h\cdot P\left(D_i\left(1\right)=h,D_i\left(0\right)\neq h\right), \end{aligned} \tag{4}$$

which gives a scaled version of the average impact of Head Start relative to alternatives among households induced to participate in Head Start by a program offer. Note that this is a variant of the Local Average Treatment Effect (LATE) concept of Imbens and Angrist (1994), where here the program "compliers" are households with $D_i\left(1\right)=h$ and $D_i\left(0\right)\neq h$.[3]

A social planner's task is to choose δ to maximize the social welfare function in (3). The first order condition for an interior optimum is:

$$g'\left(\bar{Y}\right)LATE_h=\phi_h-\phi_c S_c, \tag{5}$$

where $S_c=-\frac{\partial N_c/\partial\delta}{\partial N_h/\partial\delta}=\frac{P(D_i(1)=h,D_i(0)=c)}{P(D_i(1)=h,D_i(0)\neq h)}$ is the share of households induced to take up Head Start who would have otherwise chosen the competing program.

Efficiency dictates that a planner expand the program until the net change in social cost associated with providing the extra program slot (and reducing enrollment in competing programs) equals the dollar value of the expected test score gain relative to the next best alternative. Hence, when deciding whether to expand or contract Head Start, the evaluation problem consists of identifying $LATE_h$ and the fraction of compliers S_c drawn from the competing program.

[3]Heckman et al. (2008) extend the LATE framework to general unordered multinomial choice models (see also Kirkeboen et al., 2014). In their terminology, the LATE in (4) is the average treatment effect for those who prefer h to the optimal choice in the set $\{c,n\}$ when $Z_i=1$, but prefer the optimal choice in $\{c,n\}$ to h when $Z_i=0$.

Note that $LATE_h$ can be expressed as a weighted average of "subLATEs" measuring treatment effects on subpopulations of compliers drawn from different counterfactual alternatives:

$$\begin{aligned} LATE_h =& E\left[Y_i(h) - Y_i(c) | D_i(1) = h, D_i(0) = c\right] S_c \\ &+ E\left[Y_i(h) - Y_i(n) | D_i(1) = h, D_i(0) = n\right] (1 - S_c). \end{aligned}$$

An important lesson from (5) is that it is not necessary, at least for purposes of determining whether one is above or below optimum program scale, to separately identify these subLATEs. It is sufficient to instead identify the average impact on Head Start compliers drawn from the relevant mix of alternatives. This is natural since changes to the offer probability δ do nothing to alter the mix of compliers. Note that in the polar case where $S_c = 1$ and $\phi_h = \phi_c$ a positive $LATE_h$ would imply a "free lunch" whereby average test scores can be raised at no cost to society by shifting students out of competing programs and into Head Start. Put differently, such a result would imply competing programs should be shut down.

Structural reforms

Suppose now that the planner can alter some structural feature f of the Head Start program that households value but which has no impact on test scores. For example, f could be the quality of the transportation services provided by the Head Start center or advertising efforts targeting the Head Start-eligible population. Improving transportation services would incur additional program expenses but might draw in households who would not otherwise participate in any form of preschool. Heckman and Smith (1999) refer to such features as "program intensity variables," though they consider variables that may directly affect outcomes. Our assumption that f does not affect test scores facilitates a focus on selection effects rather than education production technology, which is beyond the scope of this paper. By shifting the composition of program participants, changes in f might (or might not) boost the program's social rate of return. We note in passing that such reforms are not purely hypothetical: 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).

To establish notation, households now value Head Start participation as $U_i(h, Z_i, f)$, where $\frac{\partial}{\partial f} U_i(h, Z_i, f) > 0$. The program feature f has no effect on the value of the other alternatives or on the joint distribution of potential outcomes. Consequently, increases in f draw children into Head Start. To reflect this, enrollments in Head Start $N_h(f)$ and the competing program $N_c(f)$ are indexed by f whenever useful. The social costs of Head Start are now written $\phi_h(f) N_h$, with $\phi_h'(f) \geq 0$.

Given our assumption of continuously distributed valuations over alternatives, one can show that:

$$\frac{d}{df}\bar{Y} = N_h'(f) \cdot MTE_h,$$

where

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

and $\overrightarrow{S}_c\left(f\right) = P\left(U_i\left(c\right) > U_i\left(n\right) | U_i(h, Z_i, f) = \max\left\{U_i(c), U_i(n)\right\}\right)$ gives the share of children on the margin of participating in Head Start who prefer the competing program to preschool non-participation. Following the terminology in Heckman et al. (2008), the marginal treatment effect MTE_h is the average effect of Head Start on test scores among households indifferent between Head Start and their next best alternative. This is essentially a marginal version of the result in (4), where integration is now over a set of children who may differ from current program participants in their mean impacts.

The planner must balance the test score effects of improvements to the program feature against the costs. An (interior) optimum ensues when:

$$g'\left(\bar{Y}\right) MTE_h = N_h \phi_h'\left(f\right) / N_h'\left(f\right) + \phi_h\left(f\right) - \phi_c \overrightarrow{S}_c\left(f\right). \tag{6}$$

Condition (6) states that the dollar value of any gain in average test scores associated with improving the feature f is to be balanced against the direct provision cost per marginal enrollee $N_h \phi_h'\left(f\right) / N_h'\left(f\right)$ of improving the program feature, plus the marginal increase in net social cost associated with expanding Head Start and contracting competing programs.[4] As in our analysis of optimal program scale, equation (6) shows that it is not necessary to separately identify the "subMTEs" that compose MTE_h to determine the socially optimal value of f. It is sufficient to identify the average causal effect of Head Start for children on the margin of participation along with the average net social cost of an additional seat in this population.

Identification using the HSIS

We have shown that assessing optimal program scale requires knowledge of $LATE_h$ and S_c, while determining optimal program features requires knowledge of MTE_h and $\overrightarrow{S_c}$. In the language of Heckman and Vytlacil (2001), $LATE_h$ and MTE_h are the policy-relevant treatment effects (PRTEs) associated with program expansion and structural reform respectively.

Our empirical analysis uses the HSIS to identify these parameters. The HSIS data are a nationally-representative random sample of Head Start applicants, and HSIS offers are distributed randomly (US DHHS, 2010). As a result, compliance patterns and treatment effects in the HSIS should mimic patterns generated by random rationing of seats among program applicants. The HSIS is therefore ideal for estimating values of $LATE_h$ and S_c *in the population of Head Start applicants.*

A complication arises if Head Start program expansion induces new households to apply to the

[4] See Appendix B for derivations of equations (5) and (6).

program, in which case the HSIS data would fail to measure a representative sample of program compliers. There are two reasons to believe the Head Start applicant pool would not change in response to a change in the offer probability δ. First, applying to Head Start is nominally free. If application costs are zero, all households with any interest in Head Start will apply regardless of the admission probability. Second, in accord with this argument, the current Head Start application rate is extremely high, which limits the scope for further selection into the applicant pool. 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 (57 percent). Fifteen percent of participants attend undersubscribed centers outside the HSIS sample, which implies that about 63 percent ($0.85 \cdot 0.57 + 0.15$) 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. While we cannot study the 15 percent of children who apply to undersubscribed Head Start centers, these calculations show that selection into the Head Start applicant pool is unlikely to be quantitatively important for our analysis.

In contrast to the analysis of program scale, the HSIS data are not ideal for assessing structural reforms. MTE_h measures the effect of Head Start for households that respond to changes in structural program features that were not directly manipulated by the HSIS experiment. Identifying program effects for these marginal households requires additional assumptions. A comparison of equations (5) and (6) reveals that MTE_h will differ from $LATE_h$ if either the "subLATEs" for compliers drawn from home care and other preschools differ from the corresponding "subMTEs" or if the weights on these subcomponents differ. Hence, to evaluate structural reforms using data from the HSIS, it is necessary to predict both how treatment effects and compliance patterns are likely to change as selection into the program is altered. In Section 7, we develop a semi-parametric econometric model that allows us to assess the effects of structural reforms using the HSIS data.

4 Data

Our core analysis sample includes 3,571 HSIS applicants with non-missing baseline characteristics and Spring 2003 test scores. Table 1 provides summary statistics for this sample. The Head Start population is disadvantaged: Column (1) shows that 40 percent of mothers in the HSIS sample are high school dropouts. Seventeen percent of mothers were teenagers at childbirth, and less than half of Head Start applicants live in two-parent households. The average applicant's household earns about 90 percent of the federal poverty line. The HSIS experiment includes two age cohorts, with 55 percent of applicants randomized at age 3 and the remaining 45 percent randomized at age 4. The bottom rows of Table 1 show statistics for two key characteristics of Head Start centers. Sixty percent of children applied to centers offering transportation services. The last row summarizes an index of Head Start center quality. This variable combines information on class size, teacher education, and other inputs to produce a composite measure of quality. It is scaled to range between

zero and one, with a mean of 0.5.[5]

Column (2) of Table 1 reports coefficients from regressions of baseline characteristics on a Head Start offer indicator to check balance in randomization. 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).[6] The results in Table 1 indicate that randomization was successful: baseline characteristics among applicants offered a slot in Head Start were similar to those denied a slot. Although children with special education status are slightly over-represented in the treatment group, this appears to be a chance imbalance – a joint test of equality of baseline characteristics fails to reject at conventional levels.

Columns (3) through (5) of Table 1 compare characteristics of children attending Head Start to those of children attending other preschool centers and no preschool. Children in non-Head Start preschools tend to be less disadvantaged than children in Head Start or no preschool, though most differences between these groups are modest. The other preschool group has a lower share of high school dropout mothers, a higher share of mothers who attended college, and higher average household income than the Head Start and no preschool groups. Children in other preschools outscore the other groups by about 0.1 standard deviations on a baseline summary index of cognitive skills (this index is described in detail below). 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.[7]

5 Experimental Impacts on Head Start Compliers

We turn now to evaluating the impact of the HSIS experiment on program compliers. Specifically, we provide non-parametric estimates of S_c and $LATE_h$, two quantities that the model of Section 3 indicated were key inputs to a cost-benefit analysis.

Substitution patterns

To investigate substitution patterns, Table 2 shows enrollment in Head Start and other preschools by age, assignment cohort, and offer status. Other preschools are a popular alternative among families denied admittance to Head Start, with this arrangement growing more popular for both offered and non-offered families by age 4. Moreover, many children denied admission manage to enroll in other Head Start centers. This is especially true for children in the three-year-old cohort,

[5] See Appendix A for further details on sample construction and variable definitions.

[6] Multiple waves of random assignment were carried out at some centers; small centers were also occasionally grouped together before random assignment. In these cases, our weights may not reflect the *ex ante* probability of a child's experimental assignment. The discussion in DHHS (2010) suggests that such cases are rare, however.

[7] Many state preschool programs enroll four-year-olds but not three-year-olds (Cascio and Schanzenbach 2013).

who could reapply to Head Start at age 4 if denied admission at age 3. By the second year of the experiment, the treatment/control difference in Head Start enrollment rates for the three-year-old cohort is 0.163 (0.657 - 0.494), and the difference in enrollment in any preschool is 0.046 (0.127 - 0.081). Evidently, the experimental offer had little effect on the probability that children in the three-year-old cohort ever enrolled in preschool. The difference in Head Start enrollment rates at age 4 is larger in the four-year-old cohort because this group did not have the opportunity to reapply to Head Start. Even so, for this cohort, the experimental offer increased the probability of enrollment in Head Start by 0.665 but only boosted the probability of enrollment in any preschool by 0.393.

We can use these figures to form estimates of the share S_c of Head Start compliers who would have otherwise enrolled in competing preschools. According to our model,

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

where Z_i can be thought of as experimental status in the HSIS experiment. Hence, we can simply scale experimental impacts on participation in competing preschools by corresponding impacts on Head Start participation to arrive at estimates of S_c.

Column (7) of Table 2 shows such estimates by cohort and year. The share of compliers drawn from other preschool centers is 0.28 for the three-year-old cohort in the first year of the experiment, and this share rises to 0.72 in the second year. The estimate of S_c for the four-year-old cohort is 0.41. These calculations show that a substantial share of experimental compliers are drawn from alternative preschool programs. S_c differs by age and applicant cohort because of the wider availability of alternative preschools at age 4, along with the opportunity for three-year-old applicants to reapply to Head Start in the second year. Pooling the three- and four-year-old cohorts in the first year of the experiment yields an estimate of S_c equal to 0.35. We use this estimate in the cost-benefit calculations below.

To evaluate the costs of Head Start, it is important to understand the characteristics of alternative preschools attended by Head Start compliers. If these programs are also financed by public subsidies, the net social cost of shifting a child into Head Start from an alternative program is likely to be small. Table 3 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 (0.154/0.606) 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.

The model in Section 3 shows that optimal policy depends on the counterfactual enrollment decisions of households induced to attend Head Start by an offer, which may differ from other

households. While these compliers cannot be directly identified in the data, the distributions of their characteristics are identified (Imbens and Rubin, 1997; Abadie, 2002). In Appendix C, we derive methods to characterize the competing programs attended by the subpopulation of compliers drawn from other preschools. Column (3) of Table 3 reports funding sources for these preschools. The results here are broadly similar to the overall means in column (2). In the absence of a Head Start offer, compliers would attend competing 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.

Finally, Table 4 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 2, 3 and 4 show that participation in alternative preschool programs is an important feature of the HSIS experiment and that these alternative programs often involve public subsidies that can generate social costs similar to Head Start.

Impacts on test scores

In assessing the impact of HSIS on test scores, we will restrict our attention to a summary index of cognitive test scores. The summary index averages Woodcock Johnson III (WJIII) test scores with PPVT scores (see Appendix A for details). This measure is normed to have mean zero and variance one among the sample of children whose applications were denied, separately by applicant 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 (US DHHS, 2010). An average of these two high-quality tests is likely to better measure cognitive skills than either test alone.

Columns (1) through (3) of Table 5 report intent-to-treat impacts of the Head Start offer on the summary index, separately by test age and assignment cohort. To increase precision, we regression-adjust these treatment/control differences using the baseline characteristics in Table 1.[8] Our results 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-

[8]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.

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 -0.01 standard deviations by kindergarten. Intent-to-treat estimates for first grade are positive, but small and statistically insignificant.

Interpretation of these intent-to-treat impacts is clouded by the preschool participation patterns in Table 2, which show substantial noncompliance with experimental assignments. Columns (4) through (6) of Table 5 report instrumental variables estimates that scale the intent-to-treat impacts by first-stage effects on Head Start attendance. The endogenous variable in these models is an indicator for attending Head Start at any time prior to the test.[9] The results can be interpreted as local average treatment effects for compliers relative to the next best alternative ($LATE_h$ in our earlier notation). The IV estimates reveal first-year impacts of 0.28 standard deviations for three-year-olds and 0.21 standard deviations for four-year-olds. Pooling three- and four-year-olds in Spring 2003 yields an estimated $LATE_h$ of 0.247 standard deviations.

Compliance for the three-year-old cohort falls after the first year as members of the control group reapply for Head Start, resulting in substantially larger standard errors for estimates in later years of the experiment. The first-grade IV estimate for the three-year-old cohort is 0.114 standard deviations, with a standard error of 0.097. Notably, the 95-percent confidence interval for first-grade impacts includes effects as large as 0.3 standard deviations. The upper bound of the confidence interval for the pooled estimate is smaller, but still substantial (0.19 standard deviations). These results show that although the longer-run impacts 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).

As a result of the imprecision of the elementary school estimates, further analysis of these outcomes is unlikely to be informative. In addition, analysis of effects beyond the first year of the experiment is complicated by the diverse patterns of program participation evident in Table 2: children in the three-year-old cohort can participate in Head Start at age 3, at age 4, both, or neither. The "correct" definition of treatment exposure is unclear and difficult to infer with a single binary instrument. Moreover, evidence from Chetty et al. (2011) suggests that immediate test score effects of early-childhood programs may predict impacts on long-run outcomes better than test score effects in other periods: classrooms that boost test scores in the short run also increase earnings in the long run, despite fadeout of test score impacts in the interim.[10] For these

[9]Our definition of treatment 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 A2 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.

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

reasons, the remainder of this paper focuses on analyzing impacts on test scores in the first year after Head Start application.

One might also be interested in the effects of Head Start on non-cognitive outcomes. Chetty et al. (2011) and Heckman et al. (2013) argue that persistent effects of early-childhood interventions on non-cognitive traits mediate long-run gains, which may explain the re-emergence of "sleeper effects" that are not evident in medium-run test scores (see also Deming, 2009). The HSIS includes short-run parent-reported measures of behavior and teacher-reported measures of teacher/student relationships. Head Start appears to have no impact on these outcomes (DHHS, 2010; Walters, forthcoming). The HSIS non-cognitive outcomes differ from those analyzed in previous studies, however, and it is unclear whether they capture the same skills.[11] In the absence of evidence that these outcomes measure traits linked to long-run outcomes, we rely on short-run test scores, which have been shown to proxy for long-run gains in a number of contexts and are often cited in policy debates.

6 Cost-Benefit Analysis

We now use our estimates to conduct a formal cost-benefit analysis of Head Start. A comparison of costs and benefits requires estimates of each term in equation (5). We obtain estimates of $LATE_h$ and S_c from the HSIS data and calibrate the other terms from estimates in the literature. The parameters underlying the cost-benefit analysis are listed in Table 6. This analysis is necessarily built on strong assumptions. The purpose of the exercise is to obtain rough estimates of Head Start's social rate of return and to determine the quantitative importance of program substitution in calculating this return.

The term $g'\left(\bar{Y}\right)$ gives the dollar value of a one standard deviation increase in test scores. A conservative valuation of test scores considers only their effects on adult earnings, as a sufficiently large dollar impact on earnings indicates the possibility for a Pareto improvement. Using data from the Tennessee STAR class size experiment, Chetty et al. (2011) estimate that a one standard deviation increase in kindergarten test scores induces a 13% increase in lifetime earnings. If test score distributions differ across populations, effects in standard deviation units may have different meanings. Statistics in Sojourner (2009) show that the standard deviation of 1st-grade scores in the STAR experiment is 87 percent of the national standard deviation. The corresponding number for the HSIS sample is approximately 81 percent. This implies that, in the HSIS, a one standard deviation increase in scores is worth slightly less than a 13 percent earnings gain. If our test score index measures the same underlying cognitive skills as the Stanford Achievement test used by Chetty et al. (2011) with the same signal to noise ratio, then we should deflate their implied earnings effects by 5 to 10 percent.[12] To be conservative, we simply assume that $g'(\bar{Y})$ equals

[11] Chetty et al. (2011) examine teacher-reported measures of initiative, and whether students annoy their classmates. Heckman et al. (2013) study 43 psychometric measures of child personality (see also Heckman, Stixrud, and Urzua, 2006). Deming (2009) looks at high school graduation, grade repetition, idleness, and learning disabilities, among other outcomes. Unlike the HSIS, none of these studies look at non-cognitive outcomes reported by parents.

[12] Sojourner (2009) shows that the standard deviation of nationally-normed percentile scores in the STAR sample

$0.1g(\bar{Y})$.

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.[13] This implies that the average child in Head Start is expected to earn 78 percent of the US average $(1 - (1 - 0.46) \cdot 0.4)$, a present value of $343,492 at age 3.4. Thus, we calculate that the marginal benefit of additional Head Start enrollment is $0.1 \cdot \$343,492 \cdot LATE_h$. Using the pooled first-year estimate of $LATE_h$ reported in Section 5, the marginal benefit of Head Start enrollment is therefore $0.1 \cdot \$343,492 \cdot 0.247 = \$8,472$.

Equation (5) shows that the net marginal social cost of Head Start enrollment depends on the relative costs of Head Start and competing programs along with the share of children drawn from other programs. Per-pupil expenditure in Head Start is approximately $\$8,000$ (DHHS, 2013). We assume a deadweight loss of taxation equal to 25 percent, which makes the gross marginal social cost of Head Start enrollment $10,000.[14] As discussed in Section 5, our estimate of the share of compliers drawn from competing programs is $S_c = 0.35$. The social cost of enrollment in other preschools depends on the extent to which competing programs are subsidized and the cost efficiency of Head Start relative to these programs. We conduct cost-benefit analyses under four assumptions: ϕ_c is either zero, 50 percent, 75 percent, or 100 percent of ϕ_h. Table 3 suggests that roughly 75 percent of competing programs are financed primarily by public subsidies, so our preferred calculation uses $\phi_c = 0.75\phi_h$.

These calculations imply that the social return to additional Head Start enrollment is positive for reasonable values of ϕ_c. With $\phi_c = 0$, costs exceed benefits: the ratio of benefits to costs is 0.85, implying a social return of -15 percent. By contrast, with $\phi_c = 0.5\phi_h$, the ratio of marginal benefit

is 24.7, 87 percent of the national standard deviation (28.3 by definition). The standard deviation of Spring 2003 PPVT scores in the HSIS is 20 percentile points, and the standard deviation of the WJIII score is 13.6 standard score points (the WJIII measure is normed to have a standard deviation of 15). Relative to the national standard deviation, these equal 70 percent and 91 percent, for a mean of 81 percent. Reliabilities for the Stanford Achievement Test, PPVT, and WJIII are 0.87, 0.94, and 0.97, so the signal-to-noise ratio is slightly larger for the Stanford test.

[13]Chetty et al. (forthcoming) 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 social return calculations are conservative in the sense that they underestimate the earnings impact of Head Start's test score gains.

[14]The exact deadweight loss will depend upon how the funds are raised. In a recent review, Saez, Slemrod, and Giertz (2012) conclude that "the marginal excess burden per dollar of federal income tax revenue raised is $0.195 for an across-the-board proportional tax increase, and $0.339 for a tax increase focused on the top 1 percent of income earners."

to cost is 1.03, so benefits exceed costs by 3 percent. With ϕ_c equal to 75 or 100 percent of ϕ_h, the benefit-cost ratio rises to 1.15 or 1.30, respectively. Our preferred estimate implies that social benefits exceed social costs by 15 percent. These results show that accounting for substition from other subsidized programs is crucial and evaluations of Head Start that neglect this substitution are likely to substantially overstate social costs. A calculation ignoring substitution from other programs yields a negative assessment of Head Start's social value. By contrast, our results suggest that expanding the scale of Head Start would produce sizable social returns at the margin.

7 Econometric Model

The above analysis reveals that accounting for the public savings associated with reductions in competing programs is pivotal for assessing whether the Head Start program pays for itself in terms of increased earnings. A separate question is whether the program can be altered in some way that makes it more effective. Short of running another experiment that manipulates a particular feature of Head Start, there is no way to answer this question nonparametrically. However, it is possible to infer how things might change in response to reforms by studying the selection patterns in the HSIS experiment and asking what would happen if these patterns held stable outside the range of observed variation.

In this section, we develop an econometric framework geared toward characterizing the substitution patterns present in the HSIS experiment and their link to test score outcomes. The goals of this analysis are twofold. First, we wish to estimate separate impacts of attendance at Head Start and competing preschools. Though not directly policy-relevant, these parameters may be of some scientific interest to researchers interested in developing new educational technologies. Second, we seek to characterize selection into the program and the relationship between the factors driving selection and outcomes. This will allow us to simulate the effects of structural reforms that change compliance patterns. To achieve identification, we exploit variation in substitution patterns across subgroups to infer the role that selection into program participation plays in generating treatment effect heterogeneity. We begin with a non-parametric subgroup analysis to illustrate the variation that drives identification of the model. We then outline the model, report estimates, and use the results to consider policy counterfactuals.

Impact heterogeneity and program substitution

Here we examine subgroups of the data, defined based upon household and site level characteristics, across which the fraction S_c of Head Start compliers who would otherwise participate in competing preschool programs varies substantially. We then ask whether the variation in S_c across these groups can explain the corresponding cross-group variation in $LATE_h$, as would be the case if each program alternative had a homogeneous effect on the level of test score outcomes.

Figure 1 provides a visual representation of this exercise, plotting group-specific IV estimates of $LATE_h$ in the first year of the experiment against corresponding estimates of S_c. The group-specific

estimates of $LATE_h$ come from an instrumental variables regression of the form:

$$Y_i = \sum_{g=1}^{\bar{G}} 1\{G_i = g\} \left(\beta_g^0 + \beta_g^h \cdot 1\{D_i = h\}\right) + X_i'\beta^x + \epsilon_i,$$

where G_i is a categorical variable indexing groups defined by the intersection of whether the household's income is above the median, mother's education, age cohort, a composite measure of preschool quality, and a measure of transportation services at the Head Start site. X_i is a vector of the same baseline covariates used in Table 5, included to increase precision. The interactions of group with Head Start attendance are instrumented by interactions of group and offer. Each parameter β_g^h captures the $LATE_h$ for group g, which we term $LATE_h^g$. Group-specific compliance share estimates S_c^g come from corresponding IV regressions using $-1\{D_i = c\}$ as the dependent variable. Groups with fewer than 100 observations are combined to form a single composite group.

Figure 1 reveals that the relationship between $LATE_h^g$ and S_c^g is sharply downward sloping. This indicates that the effects of Head Start participation are smaller among subpopulations that draw a large share of compliers from other preschools. Can this heterogeneity across groups be explained entirely by variation in the complier share S_c^g? To answer this question, we consider the null hypothesis that Head Start and competing preschools are equally effective and that the test score effects of Head Start relative to home care are constant across groups. We can write this hypothesis:

$$LATE_h^g = \tau\left(1 - S_c^g\right). \tag{7}$$

Note that this model implies that when S_c^g equals one, the $LATE_h^g$ falls to zero. This restriction also implies that all of the variation across groups in $LATE_h^g$ is attributable to heterogeneity in the counterfactual enrollment choice of Head Start compliers. To test this hypothesis we conduct optimal minimum distance estimation of (7) treating τ and the $\bar{G}$ compliance shares $\left\{S_c^1, S_c^2, ..., S_c^{\bar{G}}\right\}$ as unknown parameters and using the $2\bar{G}$ estimates $\left\{\widehat{LATE}_h^1, ..., \widehat{LATE}_h^{\bar{G}}, \hat{S}_c^1, ..., \hat{S}_c^{\bar{G}}\right\}$ as moments to match.[15] Restricted estimates of the compliance shares and the best fitting line are portrayed in Figure 1.

The restrictions in model (7) are not rejected (p-value = 0.22), suggesting that a large portion of the variation across groups in $LATE_h^g$ is attributable to variation in the care environment from which compliers are drawn. The minimum distance estimate of τ is 0.315, which implies that if all Head Start compliers were drawn from home care the $LATE_h$ would be roughly 30 percent greater than the first-year $LATE_h$ estimate of 0.247. This finding strongly suggests that the $LATE_h$ can be altered by modifying substitution patterns in the market for preschool services.

[15]This minimum distance approach bears a close connection to limited information maximum likelihood (LIML) estimation of a constant coefficient model where test scores are modeled as a function of a dummy for any preschool and the $\bar{G}$ group interactions with the offer dummy are used as excluded instruments (see Goldberger and Olkin, 1971). Our minimum distance estimator differs from LIML only because we use a cluster-robust covariance matrix to weight the reduced form moments. Note that by treating the compliance shares $\{S_{cg}\}_{g=1}^{\bar{G}}$ as unknowns, we account for the effects of estimation error in those moments. An alternate approach would be to use the errors-in-variables procedure of Deaton (1985), which Devereux (2007) shows amounts to a jackknifed instrumental variables estimator.

Table 7 corroborates this interpretation using more conventional two-stage least squares (2SLS) methods. The first row in column (1) of Table 7 displays the IV estimate of $LATE_h$ pooled across age cohorts in the first year of the experiment, 0.247. The second row uses as additional instruments linear interactions of the offer instrument and the variables used to form the groups in Figure 1. As expected given the heterogeneity in substitution patterns across these groups, adding these instruments leads to a rejection of the model's overidentifying restrictions (p-value = 0.048). Column (2) shows that treating enrollment in any preschool (Head Start or other) as the endogenous variable yields a test score impact of 0.377 standard deviations. Adding the offer interactions as instruments does little to the point estimate but this time leads to acceptance of the model at the 5 percent level, though the overidentifying restrictions are still close to being rejected (p-value= 0.070). This near rejection of the overidentifying restrictions suggests that, despite being strongly predictive of $LATE_h^g$, compliance shares may not explain all cross-group treatment effect heterogeneity. Finally, columns (3) and (4) treat Head Start and other preschools as separate endogenous variables with constant coefficients. Identification of this model requires the offer interactions. Estimates of this model show that while the instruments have substantial explanatory power for Head Start participation, they do a poor job inducing independent variation in competing preschools: the Angrist and Pischke (2009) partial F-statistic for other preschools is 1.9, well short of the standard rule of thumb of 10.0. The 2SLS estimates suggest that Head Start and other preschools have roughly similar effects, which is consistent with the restriction in (7). However, the estimates are imprecise, owing to the weak first stage for competing programs.

Selection model

The model of Section 2 indicates that determining optimal changes to program features valued by households requires knowledge of substitution patterns and their link to potential outcomes. Specifically, one would like knowledge of the schedule of marginal treatment effects MTE_h when setting policy. While it is, in principle, possible to estimate such quantities non-parametrically in research designs where excludable shifters have full support (Heckman and Vytlacil, 1999, 2005), the HSIS experiment provides us with a discrete shifter that necessitates a more structured approach.[16]

Our econometric selection model parametrizes the preferences and potential outcomes introduced in the model of Section 2. To review, program participation decisions are generated by utility maximization:

$$D_i = \underset{d \in \{h,c,n\}}{\arg\max}\, U_i\left(d, Z_i\right).$$

Normalizing the value of preschool non-participation to zero, we assume households have indirect

[16] Carneiro and Lee (2009) discuss semi-parametric estimation of marginal treatment effects in the binary treatment case. See also Doyle (2007, 2008) and Mogstad and Wiswall (2010) for recent semi-parametric approaches to estimation of marginal treatment effects.

utility over program alternatives given by:

$$\begin{aligned} U_i(h, Z_i) &= \psi_h^0 + X_i'\psi_h^x + \psi_h^z \cdot Z_i + Z_i \cdot X_i^{1\prime}\psi_h^{zx} + v_{ih}, \\ U_i(c) &= \psi_c^0 + X_i'\psi_c^x + v_{ic}, \\ U_i(n) &= 0, \end{aligned} \tag{8}$$

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 the characteristics used to form the groups in Figure 1.

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

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

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). We parameterize the correlation across alternatives 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, we allow for rich heterogeneity in program substitution patterns that can generate treatment effect heterogeneity.

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:

$$E\left[Y_i(d) | X_i, Z_i, v_{ih}, v_{ic}\right] = \theta_d^0 + X_i'\theta_d^x + \gamma_d^h v_{ih} + \gamma_d^c v_{ic}. \tag{9}$$

Assumption (9) 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, (9) actually accommodates a wide variety of data generating processes exhibiting conditional heteroscedasticity and non-normality.[17] Although it is possible to extend this model to allow the potential outcome means to depend upon higher-order

[17] For example, the conditional distribution of potential outcomes could be a location-scale mixture of normal components with density $f_d(y) = \sum_{k=1}^K \frac{1}{\sigma_{dk}(X_i)}\tilde{\phi}\left(\frac{y-\theta_{dk}^0 - X_i'\theta_{dk}^x - \gamma_{dk}^h v_{ih} - \gamma_{dk}^c v_{ic}}{\sigma_{dk}(X_i)}\right)\tilde{\pi}_{dk}(X_i)$, where $\tilde{\phi}$ is the standard normal density, $\sigma_{dk}(X_i)$ is a conditional variance function, and $\{\tilde{\pi}_{dk}(X_i)\}_{k=1}^K$ 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. It is straightforward to verify that this model obeys (9) with $\gamma_d^h = \sum_{k=1}^K \gamma_{dk}^h E\left[\tilde{\pi}_k(X_i)\right]$ and $\gamma_d^c = \sum_{k=1}^K \gamma_{dk}^c E\left[\tilde{\pi}_k(X_i)\right]$.

polynomial terms in the selection errors as in Dahl (2002), doing so would necessitate stronger instruments for estimation in the HSIS sample which is relatively small. Below we conduct some specification tests which indicate that (9) provides a reasonable approximation to mean potential outcomes.

The $\left\{\gamma_d^h, \gamma_d^c\right\}$ terms capture "essential heterogeneity," treatment effect heterogeneity that is related to selection into treatment. 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, the conditional expectation of realized outcomes can be written:

$$E\left[Y_i | X_i, Z_i, D_i = d\right] = \theta_d^0 + X_i'\theta_d^x + \gamma_d^h \lambda_d^h\left(X_i, Z_i\right) + \gamma_d^c \lambda_d^c\left(X_i, Z_i\right), \tag{10}$$

where $\lambda_d^h\left(X_i, Z_i\right) = E\left[v_{ih} | X_i, Z_i, D_i = d\right]$ and $\lambda_d^c\left(X_i, Z_i\right) = E\left[v_{ic} | X_i, Z_i, D_i = d\right]$ are multivariate generalizations of the standard inverse Mills correction term used in the Heckman (1979) selection framework. We compute the selection correction terms using formulas for truncated bivariate normal integrals derived in Tallis (1961). Appendix D provides analytical expressions.

To gain intuition regarding identification of the selection coefficients $\left\{\gamma_d^h, \gamma_d^c\right\}$, note that the control function terms $\lambda_d^h\left(X_i, Z_i\right)$ and $\lambda_d^c\left(X_i, Z_i\right)$ only vary conditional on the covariates X_i due to experimental variation in offer status Z_i. By examining how the mean test scores of individuals with treatment status d vary across households with different offer statuses, we can infer the mean potential test scores of program compliers. The control functions provide a parametric characterization of how the composition of compliers varies with the choice probabilities: these terms can be rewritten $\lambda_d^h\left(\pi_h(X_i, Z_i), \pi_c(X_i, Z_i)\right)$ and $\lambda_d^c\left(\pi_h(X_i, Z_i), \pi_c(X_i, Z_i)\right)$, where $\pi_d(X_i, Z_i) = P\left(D_i = d | X_i, Z_i\right)$. This allows us to project the effects of hypothetical policy changes that yield different choice probabilities.[18] The nonlinearities inherent in the multinomial Probit functional form aid identification by ensuring that the choice probabilities are nonseparable functions of the covariates and the Head Start offer. In practice, however, estimation of (10) requires at least one interaction in (8) of Z_i with an element of X_i to avoid severe collinearity of the control functions.

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 $\left(\hat{\lambda}_d^h\left(X_i, Z_i\right), \hat{\lambda}_d^c\left(X_i, Z_i\right)\right)$, which are included as regressors in least squares estimation of (10). When estimating the model, we renorm the covariate vector X_i to have unconditional mean zero so that the coefficients θ_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.

[18] Such representations can be had even in non-parametric models; see Dahl (2002) and Heckman and Vytlacil (2005) for discussion.

To increase precision, we also consider restrictions on the coefficients $\{\theta_d^0, \theta_d^x, \gamma_d^h, \gamma_d^c\}_{d\in\{h,c,n\}}$ 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.

Structured 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 (3) 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. The finding of a significant correlation 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 (10). 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 recently received renewed attention in the education literature (Chetty et al., 2014a; Rothstein, 2010; Kane et al., 2008). 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 into the different care alternatives. 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 modeling. 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 selection corrections remain highly collinear despite the 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.[19] This dramatically improves 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 ≥ 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 γ_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 γ_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 that 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.[20] 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. In contrast, the estimated difference between γ_c^c and γ_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.

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

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

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 estimates, wherever they exist. In this table, we refer to the subpopulation of children drawn into Head Start from competing programs as c-compliers, and the subpopulation drawn from no preschool as n-compliers. Similarly, c- and n-never takers are children that decline the Head Start offer in favor of competing programs and no preschool, respectively. The nonparametric estimates of mean potential outcomes for the complier subpopulations are computed via a generalization of the IV methods developed by Abadie (2002), which we describe in Appendix C. 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).

The model allows us to separately identify treatment effects for children drawn into Head Start from other preschools and home care. Table 11 reports some of the implied treatment effect parameters 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. More policy-relevant are the implied impacts on program compliers, which we compute by integrating over the relevant regions of X_i, v_{ih} and v_{ic}.[21]

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 estimates obtained via IV. As shown in Section 3, $LATE_h$ can be decomposed into a weighted average of "subLATEs":

$$\begin{aligned} LATE_{nh} &= E\left[Y_i\left(h\right) - Y_i\left(n\right) | D_i\left(1\right) = h, D_i\left(0\right) = n\right], \\ LATE_{ch} &= E\left[Y_i\left(h\right) - Y_i\left(c\right) | D_i\left(1\right) = h, D_i\left(0\right) = c\right]. \end{aligned}$$

These quantities give the average test score impacts of moving compliers from particular program alternatives into Head Start. Estimates of $LATE_{nh}$ are stable across specifications and indicate that the impact on program compliers 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, which would suggest the possibility of a "free lunch" associated with moving families from competing subsidized programs into Head Start.

It is worth comparing our findings with those of Feller et al. (2014), who use the the principal stratification framework of Frangakis and Rubin (2002) to estimate effects on n- and c-compliers in the HSIS. They find large effects for compliers drawn from home and negligible effects for compliers drawn from competing programs, though their point estimate of $LATE_{nh}$ is somewhat smaller

[21] For example, n-compliers have $-\psi_h(X_i, 1) < v_{ih} < -\psi_h(X_i, 0)$ and $v_{ic} < -\psi_c(X_i)$, where $\psi_h(x, z)$ and $\psi_c(x)$ are the mean utilities in equation (8).

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.

8 Evaluating Structural Reforms

We next use our structured estimates to ask how marginal costs and benefits vary with a structural program feature f, as described in Section 3. This thought experiment differs from the cost-benefit analysis in Table 6, which considers the costs and benefits of expanding the offer probability δ. An increase in f boosts the attractiveness of the program, drawing in children with weaker preferences for Head Start who would otherwise decline an offered seat. This can be viewed as a policy experiment that increases Head Start transportation services or outreach efforts to target children who would be unlikely to attend. Children indifferent between Head Start and the next best alternative at higher f may differ from compliers in the HSIS experiment. We use our structured estimates from the previous section to compute marginal treatment effects for these alternative subgroups of compliers, treating changes in f as changes to the intercept ψ_h^0 of the Head Start utility in (8).

Changes to program features that make Head Start more attractive may also increase the direct costs of the program. The term $N_h\phi_h'(f)/N_h'(f)$ in equation (6) captures this effect. This term can be written $\eta \cdot \phi_h$, where $\eta = d\ln\phi_h/d\ln N_h$ 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 η. We start with the extreme case where $\eta = 0$, which allows us to characterize costs and benefits associated with reforms that draw in children on the margin without changing the per-capita cost of the program. We then consider how the cost-benefit calculus changes when $\eta > 0$.

Figure 2 summarizes marginal costs and benefits as a function of f. 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 figure shows that as the Head Start attendance rate rises, the benefit for children on the margin increases. This reflects the pattern of selection described in Table 9: children with weaker tastes for Head Start receive larger gains from Head Start attendance, so the benefit for marginal children increases with the scale of the program.

Figure 2 also displays several marginal cost curves corresponding to various values of ϕ_c. When $\phi_c = 0$, marginal costs are constant at \$10,000 per child, which exceeds marginal benefits for most values of f. An optimal solution here would be to shut the Head Start program down. For more plausible values of ϕ_c, marginal benefits exceed marginal costs for most values of f, and

the difference between benefits and costs increases as the program expands. This implies that the marginal condition in Section 3 does not characterize the social optimum; since benefits are convex, our estimates imply that the social optimum is a corner solution with all children attending preschool.

Finally, the red dashed line in Figure 2 shows marginal costs when $\phi_c = 0.75\phi_h$ and $\eta = 0.25$.[22] This scenario implies steeply rising marginal costs of Head Start provision: an increase in f that doubles enrollment raises per-capita costs by 25 percent. The marginal cost line crosses the marginal benefits line from below slightly above the point $f = 0$. Hence, if $\eta = 0.25$, the program feature f is roughly at its optimal level in the current incarnation of Head Start. For larger values of η, expanding the program by increasing f fails a cost-benefit test, while for smaller values of η, the social benefits of increasing f exceed the costs. 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 benefits to attracting such students will exceed the costs unless per-capita program costs increase rapidly with enrollment. These results illustrate the importance of selection and effect heterogeneity in assessments of program reforms. Our estimates suggest that structural reforms targeting children who are currently unlikely to attend Head Start and children who are likely to be drawn from nonparticipation will generate larger effects than reforms that simply create more seats. Our results also echo other recent studies finding increasing returns to early-childhood investments, though the mechanism generating increasing returns in these studies is typically dynamic complementarity in human capital investments rather than selection and effect heterogeneity (see, e.g., Cunha et al., 2010).

9 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 also find evidence that the program's social rate of return can be boosted by reforms that draw in additional households with weak unobserved tastes for the program, 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 populations. One would

[22] For this case, marginal costs are obtained by solving the differential equation $\phi_h'(f) = \eta \cdot \phi(f) \cdot (N_h'(f)/N_h(f))$, with the initial condition $\phi_h(0) = \$10,000$. This yields the solution $\phi_h(f) = \$10,000 \cdot \exp\left(\eta \left(\ln N_h(f) - \ln N_h(0)\right)\right)$.

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 study is constrained by the design of the HSIS experiment, which sampled Head Start applicants. This population is relevant for considering small changes in the rationing scheme used to allocate slots in Head Start, but very large changes could lead to equilibrium effects that lead the composition of applicants to change. Along the same lines, large changes in program features could alter the test score impacts of Head Start; for example, implementing recent proposals for universal preschool could generate a shortage of qualified teachers (Rothstein, forthcoming). Finally, our cost-benefit calculations rely on estimates of the link between test score effects and earnings gains in the Tennessee STAR class size experiment (Chetty et al., 2011). 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.

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 such settings, provided that data are available on enrollment in competing programs. An important question for future research is whether similar exercises can be conducted in the absence of detailed microdata on substitute program enrollments.

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

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

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

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

7. 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.

8. 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).

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

10. 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).

11. 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).

12. 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).

13. Chetty, R., Hendren, N., Kline, P., & Saez, E. (forthcoming). "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." Quarterly Journal of Economics.

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

15. 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.

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

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

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

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

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

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

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

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

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

25. 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.

26. 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).

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

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

29. 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.

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

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

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

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

34. 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).

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

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

37. 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.

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

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

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

41. 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).

42. 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.

43. 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).

44. 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).

45. 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).

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

47. 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.

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

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

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

51. 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).

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

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

54. 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`.

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

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

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

58. 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).

59. 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).

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

61. 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).

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

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

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

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

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

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

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

69. 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).

70. 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.

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

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

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

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

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

76. 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.

77. 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 .

78. 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 .

79. 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.

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

81. 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 (2014) 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.

Child 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 arrangments 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

Optimal program scale

This appendix derives the conditions for optimal program scale and program features in equations (5) and (6). From equation (3), the first-order condition for the optimal value of δ is given by

$$g'\left(\bar{Y}\right)\frac{d\bar{Y}}{d\delta} = \phi_h \frac{\partial N_h}{\partial \delta} + \phi_c \frac{\partial N_c}{\partial \delta}.$$

Using equation (2), we can re-write average test scores as

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

Therefore, we have

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

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

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

which is equation (4).

Next, we can write

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

so

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

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

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

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

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

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

$$g'\left(\bar{Y}\right) \cdot LATE_h \cdot P\left(D_i(1) = h, D_i(0) \neq h\right) =$$
$$\phi_h \cdot P\left(D_i(1) = h, D_i(0) \neq h\right) - \phi_c \cdot P\left(D_i(1) = h, D_i(0) = c\right).$$

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

$$g'\left(\bar{Y}\right) \cdot LATE_h = \phi_h - \phi_c \cdot \left(\frac{P\left(D_i(1) = h, D_i(0) = c\right)}{P\left(D_i(1) = h, D_i(0) \neq h\right)}\right)$$

$$\implies g'\left(\bar{Y}\right) \cdot LATE_h = \phi_h - \phi_c \cdot S_c,$$

which is equation (5).

Optimal program features

From equation (3), the first-order condition for the optimal value of f is

$$g'\left(\bar{Y}\right)\frac{d\bar{Y}}{df} = N_h\phi_h'(f) + \phi_h N_h'(f) + \phi_c N_c'(f).$$

We can rewrite equation (2) as

$$\bar{Y} = E\left[Y_i(h) \cdot 1\left\{U_i(h, Z_i, f) > \max\left\{U_i(c), U_i(n)\right\}\right\}\right]$$

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

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

Using Leibniz's rule for differentiation under the integral sign, we have

$$\frac{d\bar{Y}}{df} = E\left[Y_i(h) \cdot 1\left\{U_i(h, Z_i, f) = \max\left\{U_i(c), U_i(n)\right\}\right\}\right]$$

$$-E\left[Y_i\left(c\right) \cdot 1\left\{U_i(c) = U_i(h, Z_i, f)\right\} \cdot 1\left\{U_i(c) > U_i(n)\right\}\right]$$

$$-E\left[Y_i(n) \cdot 1\left\{U_i(n) = U_i(h, Z_i, f)\right\} \cdot 1\left\{U_i(n) > U_i(c)\right\}\right],$$

which can be rewritten

$$\frac{d\bar{Y}}{df} = E\left[\left(Y_i(h) - Y_i\left(c\right)\right) \cdot 1\left\{U_i(h, Z_i, f) = U_i(c)\right\} \cdot 1\left\{U_i(c) > U_i(n)\right\}\right]$$

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

$$=E\left[Y_i(h)-Y_i(c)|U_i(h,Z_i,f)=U_i(c),U_i(c)>U_i(n)\right]P\left(U_i(h,Z_i,f)=U_i(c),U_i(c)>U_i(n)\right)$$

$$+E\left[Y_i(h)-Y_i(n)|U_i(h,Z_i,f)=U_i(n),U_i(n)>U_i(c)\right]P\left(U_i(h,Z_i,f)=U_i(n),U_i(n)>U_i(c)\right)$$

$$=MTE_h\cdot P\left(U_i(h,Z_i,f)=\max\left\{U_i(c),U_i(n)\right\}\right).$$

The share of households attending the computing program is

$$N_c(f)=E\left[1\left\{U_i(c)>\max\left\{U_i(h,Z_i,f),U_i(n)\right\}\right\}\right]$$

$$=E\left[1\left\{U_i(c)>U_i(h,Z_i,f)\right\}\cdot 1\left\{U_i(c)>U_i(n)\right\}\right].$$

Again using Leibniz's rule, we have

$$N_c'(f)=-E\left[1\left\{U_i(c)=U_i(h,Z_i,f)\right\}\cdot 1\left\{U_i(c)>U_i(n)\right\}\right]$$

$$=-E\left[1\left\{U_i(h,Z_i,f)=\max\left\{U_i(c),U_i(n)\right\},U_i(c)>U_i(n)\right\}\right]$$

$$=-P\left(U_i(h,Z_i,f)=\max\left\{U_i(c),U_i(n)\right\}\right)\cdot\overrightarrow{S}_c(f).$$

A similar calculation shows that

$$N_h'(f)=P\left(U_i(h,Z_i,f)=\max\left\{U_i(c),U_i(n)\right\}\right).$$

The first-order condition is therefore

$$g'\left(\bar{Y}\right)MTE_hN_h'(f)=N_h(f)\phi_h'(f)+\phi_hN_h'(f)-\phi_c\overrightarrow{S_c}(f).$$

Dividing by $N_h'(f)$, we have

$$g'\left(\bar{Y}\right)MTE_h=N_h(f)\phi_h'(f)/N_h'(f)-\phi_c\overrightarrow{S}_c(f),$$

which is equation (6).

Appendix C: 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. This simplifies to

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

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,

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

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

Note that κ_c is the population coefficient from an instrumental variables regression of $g(Y_i, X_i) \cdot 1\left\{D_i = c\right\}$ 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 Table 3, 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 Table 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 D: 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 (8) 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

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

where

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

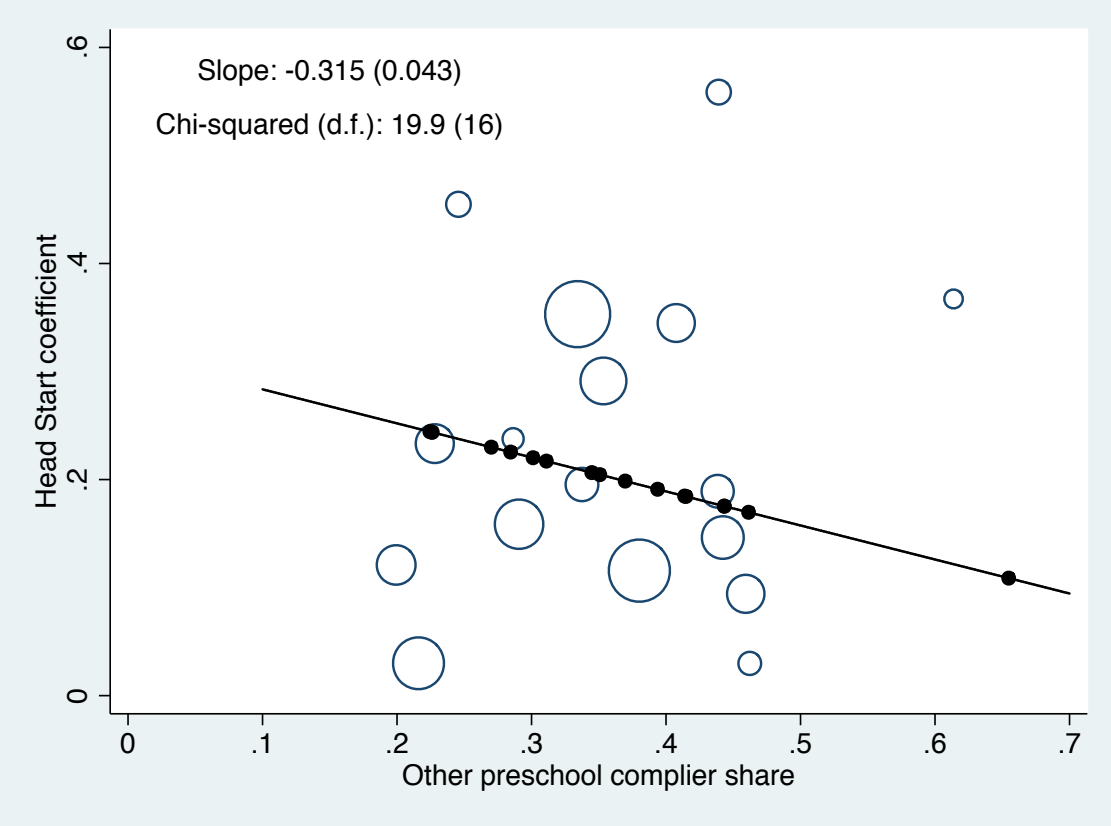
Defining $\lambda_d^k\left(X_i, Z_i\right) \equiv E\left[v_{ik} | X_i, Z_i, D_i = d\right]$, this implies that we can write

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

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

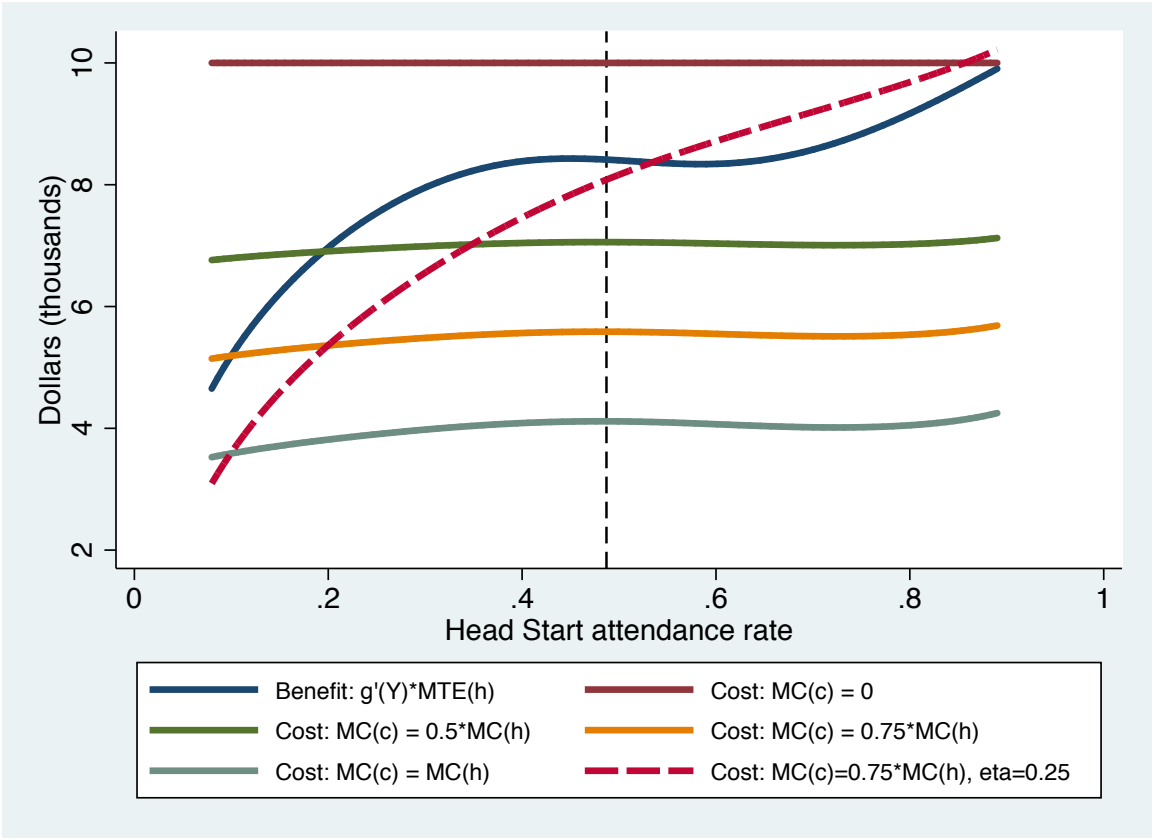
$$\begin{aligned}
\lambda_c^h(X_i, Z_i) = & -\Lambda\left(\psi_c(X_i), \tfrac{\psi_c(X_i)-\psi_h(X_i,Z_i)}{\sqrt{2(1-\rho(X_i))}}; \sqrt{\tfrac{1-\rho(X_i)}{2}}\right) \\
& + \sqrt{2\left(1-\rho(X_i)\right)} \cdot \Lambda\left(\tfrac{\psi_c(X_i)-\psi_h(X_i,Z_i)}{\sqrt{2(1-\rho(X_i))}}, \psi_c(X_i); \sqrt{\tfrac{1-\rho(X_i)}{2}}\right),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

Figure 1: Complier Shares and Head Start Effects



Notes: This table plots IV coefficients against other-preschool complier shares in strata defined by transportation, age 4, above-median income, above-median center quality, and whether a child's mother is a high school dropout. Strata with fewer than 100 observations are combined into a single group. The line is the slope coefficient from a classical minimum distance model imposing that all points are on a line through (1,0). Black points are fitted values using estimates of the complier shares from this model. The chi-squared statistic is the minimized criterion function from the minimum distance model. Circles are proportional to the reciprocal of the squared standard error of the Head Start coefficient.

Figure 2: Marginal Costs and Benefits of Head Start



Notes: This figure plots marginal costs and benefits of additional Head Start enrollment 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*, and the curves show costs and benefits for children on the margin of Head Start attendance at each *f*. The black line corresponds to the sample Head Start attendance rate (*f*=0). MTEs are computed using the restricted two-step estimates of the structural model. The marginal cost of Head Start enrollment is assumed to be $10,000 per child at *f*=0.

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: Preschool Choices by Year, Cohort, and Offer Status

		Offered			Not offered			Other center
		Head Start	Other centers	No preschool	Head Start	Other centers	No preschool	complier share
Time period	Cohort	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Age 3	3-year-olds	0.851	0.058	0.092	0.147	0.256	0.597	0.282
Age 4	3-year-olds	0.657	0.262	0.081	0.494	0.379	0.127	0.719
	4-year-olds	0.787	0.114	0.099	0.122	0.386	0.492	0.410

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) gives an estimate of the share of experimental 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 3: 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 children who would be induced to attend Head Start by an experimental offer.

Table 4: Characteristics of Head Start and Competing Preschool Centers

	Head Start (1)	Other centers (2)	Other centers attended by *c* -> *h* compliers (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 children who would be induced to attend Head Start by an experimental offer.

Table 5: Experimental Impacts on Test Scores

		Intent-to-treat			Instrumental variables		
		Three-year-olds	Four-year-olds	Pooled	Three-year-olds	Four-year-olds	Pooled
Time period		(1)	(2)	(3)	(4)	(5)	(6)
Age 3		0.194	-	-	0.278	-	-
		(0.029)			(0.041)		
	N	1970			1970		
Age 4		0.089	0.141	0.114	0.249	0.213	0.227
		(0.029)	(0.029)	(0.020)	(0.080)	(0.044)	(0.040)
	N	1760	1601	3361	1760	1601	3361
Kindergarten		-0.008	-0.021	-0.012	-0.023	-0.031	-0.023
		(0.031)	(0.036)	(0.024)	(0.084)	(0.053)	(0.047)
	N	1659	1432	3091	1659	1432	3091
1st grade		0.039	0.053	0.045	0.114	0.079	0.091
		(0.033)	(0.038)	(0.025)	(0.097)	(0.057)	(0.052)
	N	1599	1405	3004	1599	1405	3004

Notes: This table reports intent-to-treat and instrumental variables estimates of effects on a summary index of test scores. Columns (1)-(3) report coefficients from regressions of test scores on an indicator for assignment to Head Start. Columns (4)-(6) use the assignment indicator as an indicator for Head Start attendance, defined as an indicator equal to one if a child attended Head Start at any time prior to the test. 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 at the center level.

Table 6: Benefits and Costs of Head Start

Parameter	Description	Value	Source
(1)	(2)	(3)	(4)
$g'(Y)$	Effect of a 1 SD increase in test scores on earnings	$0.1*g(Y) = 0.1*w$	Chetty et al. 2011
w_{avg}	US average present discounted value of lifetime earnings at age 3.4	\$438,000	Chetty et al. 2011 with 3% discount rate
w_{parent}/w_{avg}	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
w_{hs}	Average present discounted value of lifetime earnings for Head Start applicants	\$343,392	$[1 - (1 - w_{parent}/w_{avg})*IGE]*w_{avg}$
$g'(Y_{hs})$	Effect of a 1 SD increase in test scores on earnings of Head Start applicants	\$34,339	$0.1*w_{hs}$
$LATE_h$	Local Average Treatment Effect	0.247	HSIS
MB	Marginal social benefit of Head Start enrollment	\$8,482	$g'(Y_{hs})*LATE_h$
ϕ_h	Marginal cost of Head Start enrollment	\$10,000	Head Start program facts with 25% DWL of taxation
ϕ_c	Marginal cost of enrollment at other preschools	\$0, \$5,000, \$7,500, or \$10,000	Assumption
S_c	Share of Head Start population drawn from other preschools	0.35	HSIS
MC	Net marginal social cost of Head Start enrollment	\$10,000 \$8,250, \$7,375, or \$6,500	$\phi_h - \phi_c*S_c$
MB/MC	Benefit/cost ratio	0.85, 1.03, 1.14, or 1.29	-

Notes: This table reports results of a cost/benefit calculation for Head Start. Estimated parameter values are obtained from the sources listed in column (4).

Table 7: Two-stage Least Squares Estimates of Preschool Effects

	Single endogenous variable		Two endogenous variables	
	Head Start	Any preschool	Head Start	Other centers
Model	(1)	(2)	(3)	(4)
Just-identified	0.247	0.377	-	-
	(0.031)	(0.048)		
Overidentified	0.238	0.361	0.360	0.358
	(0.030)	(0.046)	(0.148)	(0.419)
First-stage *F*	421.7	110.2	19.4	1.9
Overid. *p*-value	0.048	0.070	0.038	

Notes: This table reports two-stage least squares estimates of the effects of Head Start and other preschool centers in Spring 2003. Columns (1) and (2) show estimates of models treating either Head Start or any preschool as the endogenous variable. Columns (3) and (4) show estimates of a model treating Head Start and other preschools as separate endogenous variables. Just-identified models instrument 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 the main effects of the interacting variables and the baseline covariates listed in the notes to Table 5. Standard errors are clustered at the center level. *F*-statistics are Angrist/Pischke (2009) partial *F*'s.

Table 8: Multinomial Probit Estimates

	Head Start utility		Other center utility	Arctanh ρ
	Main effect	Offer interaction		
	(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. *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 the baseline covariates listed in the notes to Table 5. Likelihood contributions are weighted by the reciprocal of the probability of experimental assignments. 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	0.214	0.722	0.437	0.456	0.473
		(0.037)	(0.022)	(0.198)	(0.120)	(0.116)	(0.110)
$\theta_c^0 - \theta_n^0$	Effect of other preschools relative to no preschool	0.262	0.149	0.406	0.172	0.372	0.473
		(0.052)	(0.033)	(0.596)	(0.274)	(0.234)	(0.110)
γ_h^h	Coefficient on Head Start taste in Head Start outcome equation	-	-	-0.147	-0.150	-0.132	-0.137
				(0.051)	(0.051)	(0.046)	(0.044)
γ_c^h	Coefficient on Head Start taste in other preschool outcome equation		-	-0.008	-0.029	-0.132	-0.137
				(0.182)	(0.082)	(0.046)	(0.044)
γ_n^h	Coefficient on Head Start taste in no preschool outcome equation	-	-	0.084	-0.004	0.003	0.008
				(0.077)	(0.056)	(0.055)	(0.054)
γ_h^c	Coefficient on other preschool taste in Head Start outcome equation	-	-	0.030	-0.073	-0.037	-0.109
				(0.341)	(0.329)	(0.151)	(0.032)
γ_c^c	Coefficient on other preschool taste in other preschool outcome equation	-	-	0.163	0.123	-0.037	-0.109
				(0.509)	(0.186)	(0.151)	(0.032)
γ_n^c	Coefficient on other preschool taste in no preschool outcome equation	-	-	-0.754	-0.242	-0.286	-0.319
				(0.349)	(0.220)	(0.208)	(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: Mean Potential Outcomes for Subpopulations

	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.

Table A1: 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 A2: 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 withmissing values for either characteristics of the center of random assignment or the center attended.