# Online Appendix to "Evaluating Public Programs with Close Substitutes: The Case of Head Start"

Patrick Kline
UC Berkeley/NBER

Christopher Walters
UC Berkeley/NBER

April 2016

## Contents

# 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.

## A.1 Test scores

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

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

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

## A.2 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.

### A.3 Preschool attendance

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

### A.4 Preschool characteristics

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

### A.5 Weights

The probability of assignment to Head Start differed across experimental sites. The HSIS data includes several weight variables designed to account for these differences. These weights also include a factor that adjusts for differences in the probability that Head Start centers themselves were sampled (Puma et al. 2010). This weighting can be used to estimate the average effect of Head Start participation in the US, rather than the average effect in the sample; these parameters may differ if effects differ across sites in a manner related to sampling probabilities. Probabilities of sampling differed widely across centers, however, leading to very large differences in weights across children and decreasing precision. Instead of using the HSIS weights, we constructed inverse probability weights based on the fraction of applicants at each site offered Head Start. The discussion

in Puma et al. (2010) suggests that the numbers of treated and control students at each site were specified in advance, implying that this fraction correctly measures the *ex ante* probability that a child is assigned to the treatment group. Results using other weighting schemes were similar, but less precise.

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

## Appendix B: Identification of Complier Characteristics

This appendix extends results from Imbens and Rubin (1997) and Abadie (2002) to show identification of population shares, characteristics and marginal potential outcome distributions for subpopulations of compliers drawn from other preschools and no preschool. Under the monotonicity restriction (1), we have

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

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

$$=S_c.$$

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

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

$$\kappa_c\equiv\frac{E\left[g(Y_i,X_i)\cdot1\left\{D_i=c\right\}|Z_i=1\right]-E\left[g\left(Y_i,X_i\right))\cdot1\left\{D_i=c\right\}|Z_i=0\right]}{E\left[1\left\{D_i=c\right\}|Z_i=1\right]-E\left[1\left\{D_i=c\right\}|Z_i=0\right]}.$$

The numerator can be written

$$E\left[g\left(Y_i(D_i(1)),X_i\right)\cdot1\left\{D_i(1)=c\right\}\right]-E\left[g(Y_i(D_i(0)),X_i)\cdot1\left\{D_i(0)=c\right\}\right],$$

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

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

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

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

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

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

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

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

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

which completes the proof.

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

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

## Appendix C: Derivation of Marginal Value of Public Funds

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

### C.1 Program Scale

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

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

The test score for child $i$ can be written

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

so

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

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

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

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

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

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

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

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

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

which is equation (6). It follows that

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

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

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

The probability of Head Start participation is

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

which implies

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

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

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

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

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

which is equation (7).

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

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

By plugging in expressions for these derivatives we obtain

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

which is equation (8).

### C.2 Rationed Substitutes

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

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

and preschool enrollment choices are defined by

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

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

$$P(D_i = c) = \delta_h\delta_c\pi_c(1,1) + \delta_h(1-\delta_c)\pi_c(1,0) + (1-\delta_h)\delta_c\pi_c(0,1) + (1-\delta_h)(1-\delta_c)\pi_c(0,0). \quad \text{(A1)}$$

We assume that competing preschools adjust $\delta_c$ so that $dP(D_i = c)/d\delta_h = 0$. Totally differentiating equation (A1) with respect to $\delta_h$ yields

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

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

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

Average test scores are given by

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

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

so

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

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

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

which can be rewritten

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

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

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

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

Here the local average treatment effects are defined as

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

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

This can be further simplified to

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

The effect of an increase in $\delta_h$ on the government's budget is

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

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

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

Then the marginal value of public funds is

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

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

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

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

which is equation (9).

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

### C.3 Structural Reforms

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

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

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

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

We can write mean test scores as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The corresponding effects on choice probabilities are

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

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

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

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

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

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

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

which is equation (12).

### C.4 Valuing test score impacts

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

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

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

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

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

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

Define the excess expenditure function:

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

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

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

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

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

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

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

$$\begin{aligned} CV(y_0, y_1) &\approx (1-\tau) p l_c(w(y_0), u_0)(y_1 - y_0) \\ &= (1-\tau) p l(w(y_0), b)(y_1 - y_0). \end{aligned} \tag{A2}$$

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

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

which can be written:

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

where $\epsilon \equiv \frac{w(y_0)}{l(w(y_0),b)} \frac{\partial}{\partial \omega} l\left(w\left(y_0\right), b\right)$ gives the uncompensated elasticity of labor supply. Relative to (A2), this expression has an extra term $(1+\epsilon)$ that reflects how the child adjusts her lifetime labor supply in response to the increase in her after-tax wage. By the envelope theorem, these behavioral changes (when they are small) do not yield additional utility.

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

# Appendix D: Empirical Cost Benefit Analysis

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

## D.1 Representativeness of the HSIS data

The HSIS data are a nationally-representative random sample of Head Start applicants, and HSIS offers are distributed randomly (Puma et al. 2010). The HSIS is therefore ideal for estimating values of $LATE_h$ and $S_c$ *in the population of Head Start applicants.*[1] Fortunately, the current Head Start application rate is high, which limits the scope for selection into the applicant pool that might change with program scale. Currie (2006) reports that two-thirds of eligible children participated in Head Start in 2000. This is higher than the Head Start participation rate in the HSIS sample (49 percent). However, fifteen percent of participants attend undersubscribed centers outside the HSIS sample, which implies that about 57 percent $(0.85 \cdot 0.49 + 0.15)$ of all applicants participate in Head Start (Puma et al. 2010). For this to be consistent with a participation rate of two-thirds among eligible households, virtually all eligible households must apply. Therefore, selection into the Head Start applicant pool is unlikely to be quantitatively important for our analysis.

## D.2 Program benefits

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

Appendix Table A.IV summarizes several studies that compare test score and earnings impacts for the same intervention. The most closely related study is by Chetty et al. (2011), an analysis of the Tennessee STAR class size experiment. Chetty et al. (2011, p.7 online appendix) show that a one standard deviation increase in kindergarten test scores induced by an experimental change in classroom quality yields a 13.1 percent increase in earnings at age 27.[2] The STAR results also suggest that immediate test score effects of early-childhood programs predict earnings gains better

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

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

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

The STAR classroom quality estimate of 13.1 percent is smaller than a corresponding OLS estimate controlling for rich family characteristics in the STAR sample (18 percent), and comparable to estimates from Chetty, Friedman and Rockoff (2014b) linking test score and earnings impacts for teacher value-added (10.3 percent for value-added, 12 percent for OLS with controls). The Chetty, Friedman and Rockoff (2014b) findings also replicate the pattern of long-run earnings impacts coupled with fadeout of medium-run test score effects. In an analysis of the Perry Preschool Project, Heckman et al. (2010b) estimate larger ratios of earnings per standard deviation of test scores (24 to 29 percent). Sibling fixed effects estimates from studies of Head Start by Currie and Thomas (1995) and Garces, Thomas and Currie (2002) suggest much larger ratios, though the earnings estimates are also very statistically imprecise. To be conservative, our baseline calibrations assume an earnings impact of 10 percent per standard deviation of earnings, which is at the bottom of the range of estimates reported in Table A.IV.[3]

Calculating percentage changes in earnings requires a prediction of average earnings in the HSIS population. Chetty et al. (2011) calculate that the average present discounted value of earnings in the United States is approximately $522,000 at age 12 in 2010 dollars. Using a 3-percent discount rate, this yields a present discounted value of $438,000 at age 3.4 (the average age of applicants in the HSIS). Children who participate in Head Start are disadvantaged and therefore likely to earn less than the US average. The average household participating in Head Start earned 46 percent of the US average in 2013 (US DHHS, 2013; Noss, 2014). Lee and Solon (2009) find an average intergenerational income elasticity in the United States of roughly 0.4, implying that the average child in Head Start is expected to earn 78 percent of the US average $(1 - (1 - 0.46) \cdot 0.4)$.[4] These calculations yield a present value of earnings $\bar{e}$ equal to $\$343,492$ at age 3.4.

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

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

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

### D.3 Program costs

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

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

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

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

## Appendix E: Interacted Two-stage Least Squares

This Appendix investigates the use of the interacted two-stage least squares approach described in Section VII to estimate models treating both Head Start and other preschools as endogenous variables. Suppose there is a single binary covariate $X_i \in \{0, 1\}$. Under the assumptions described in Section IV, covariate-specific instrumental variables coefficients give local average treatment effects:

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

Furthermore, we have

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

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

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

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

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

The solution to this system is

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

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

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

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

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

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

which can be written

$$\begin{aligned} LATE_{nh} = &\frac{S_c(0)S_c(1)}{S_c(0)-S_c(1)} \cdot (LATE_{ch}(1) - LATE_{ch}(0)) \\ &+ (w_n(1)LATE_{nh}(1) + (1 - w_n(1))LATE_{nh}(0)), \end{aligned} \tag{A3}$$

where

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

Equation (A3) shows that the interacted 2SLS strategy yields a Head Start coefficient equal to a weighted average of the subLATEs $LATE_{nh}(x)$, plus a term that depends on heterogeneity in $LATE_{ch}(x)$. If there is heterogeneity in this other subLATE, this strategy does not recover the causal effect of $h$ relative to $n$ for any well-defined subpopulation. This result is a special case of the results in Kirkboen, Leuven and Mogstad (forthcoming) and Hull (2015), who show that 2SLS does not generally recover causal effects in models with multiple endogenous variables.

# Appendix F: Selection Model

## F.1 Control Functions

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

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

which can be re-written

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

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

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

where

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

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

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

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

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

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

## F.2 Identification

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

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

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

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

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

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

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

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

Solving these equations for the selection coefficients yields

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

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

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

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

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

### F.3 Estimating SubLATEs

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

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

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

We can rewrite these conditions

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

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

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

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

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

where

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

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

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

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

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

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

where

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

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

## F.4 Specification tests

### Testing for underidentification

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

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

Equivalently, $\gamma_{dh}$ and $\gamma_{dc}$ are not identifed if

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

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

$$\hat{\lambda}_h(X_i,1,d) - \hat{\lambda}_h(X_i,0,d) = \sum_{k=0}^{3} q_{dk} \left(\hat{\lambda}_c(X_i,1,d) - \hat{\lambda}_c(X_i,0,d)\right)^k + \eta_{id} \tag{A4}$$

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

To account for estimation error in the first-step multinomial probit parameters we conduct inference via the nonparametric bootstrap. Let $\hat{q}_d = (\hat{q}_{d0}, \hat{q}_{d2}, \hat{q}_{d3})'$ denote full-sample estimates from equation (A4) and let $\hat{q}_d^b$ denote corresponding estimates in bootstrap sample $b$. We form the test statistic

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

where

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

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

**Testing additive separability**

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

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

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

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

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

Table VII reports $p$-values for this test.

**Testing model fit**

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

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

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

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

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

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

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

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

where

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

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

## Appendix G: Site Group Fixed Effects

This appendix describes methods for incorporating experimental site group fixed effects into our two-step control function estimation procedure. These methods allow us to leverage cross-site variation while reducing the dimension of heterogeneity across sites, eliminating an incidental parameters problem that would arise with a full set of site fixed effects. Our approach is similar in spirit to that of Bonhomme and Manresa (2015), who develop methods that account for grouped patterns of heterogeneity in linear panel data models. Saggio (2012) extends the group fixed effects approach to panel binary choice models. In the translation from panel data to our multi-site experimental setting, sites play the role of cross-sectional units and experimental subjects play the role of time periods.

### G.1 Model

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

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

with

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

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

### G.2 Estimation

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

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

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

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

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

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

### G.3 Implementation

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

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

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

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

Here the $S$ in the second term captures parameters corresponding to group assignments for the $S$ sites, while the $4G$ captures the estimated group-specific parameters. The final number of groups

is chosen to minimize $BIC(G)$. As shown in Appendix Table A.VI, the BIC selects 7 groups when the model includes no other covariates and 6 groups when the model includes our full set of baseline covariates.

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

## Additional Appendix References


1. Abadie, Alberto, "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variables Models," *Journal of the American Statistical Association*, 97 (2002), 284-292.

2. Ackerberg, Daniel, and Paul Devereux, "Improved JIVE Estimators for Overidentified Linear Models With and Without Heteroskedasticity," *The Review of Economics and Statistics*, 91 (2009), 351-362.

3. Angrist, Joshua, Guido Imbens, and Alan Krueger, "Jackknife Instrumental Variables Estimation," NBER Working Paper No. 172, 1995.

4. Ashenfelter, Orley, Kirk Doran, and Bruce Schaller, "A Shred of Credible Evidence on the Long-run Elasticity of Labor Supply," *Economica*, 77 (2010), 637-650.

5. Blundell, Richard, Luigi Pistaferri, and Itay Saporta-Eksten, "Consumption Inequality and Family Labor Supply," *American Economic Review*, 106 (2016), 387-435.

6. Cascio, Elizabeth, and Douglas Staiger, "Knowledge, Tests, and Fadeout in Educational Interventions," NBER Working Paper No. 18038, 2012.

7. Chao, John, Jerry Hausman, Whitney Newey, Norman Swanson, and Tiemen Woutersen, "Testing Overidentifying Restrictions With Many Instruments and Heteroskedasticity," *Journal of Econometrics*, 178 (2014), 15-21.

8. Currie, Janet, "The Take-up of Social Benefits," in *Poverty, the Distribution of Income, and Public Policy,* Alan Auerbach, David Card, and John Quigley, eds. (New York, NY: The Russell Sage Foundation, 2006).

9. Currie, Janet, and Duncan Thomas, "Early Test Scores, Socioeconomic Status and Future Outcomes," NBER Working Paper No. 6943, 1999.

10. Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50 (1982), 1029-1054.

11. Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz, "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program," *Quantitative Economics*, 1 (2010b), 1-46.

12. Heckman, James, Jora Stixrud, and Sergio Urzua, "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24 (2006), 411-482.

13. Imbens, Guido, and Donald Rubin, "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies*, 64 (1997), 555-574.

14. Krueger, Alan, "Economic Considerations and Class Size," *Economic Journal*, 113 (2003), F34-F63.

15. Lindqvist, Erik, and Roine Vestman, "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment," *American Economic Journal: Applied Economics*, 3 (2011), 101-128.

16. Murnane, Richard, John Willet, and Frank Levy, "The Growing Importance of Cognitive Skills in Wage Determination," *The Review of Economics and Statistics*, 77 (1995), 251-266.

17. Sojourner, Aaron, "Inference on Peer Effects With Missing Peer Data: Evidence from Project STAR," *Economic Journal*, 123 (2013), 574-605.

18. Tallis, G. M., "The Moment Generating Function of the Truncated Multi-normal Distribution," *Journal of the Royal Statistical Society*, 23 (1961), 223-229.

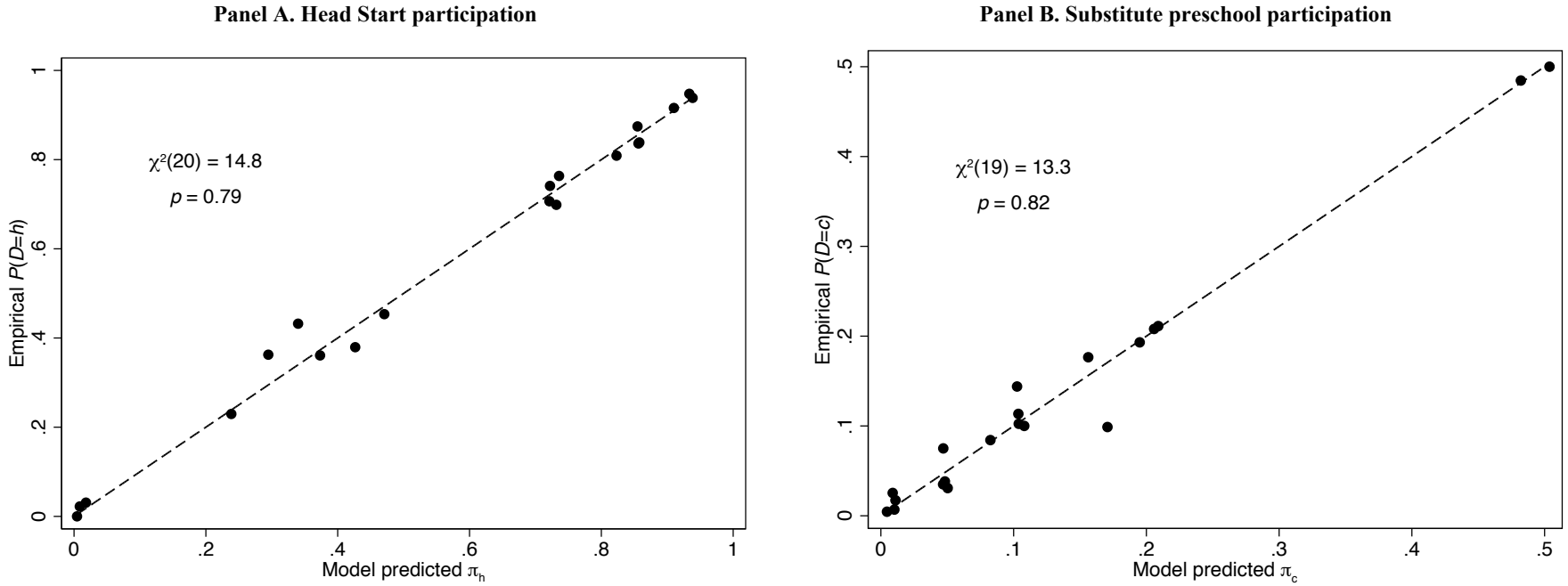


Figure A.I. Multinomial Probit Model Fit

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

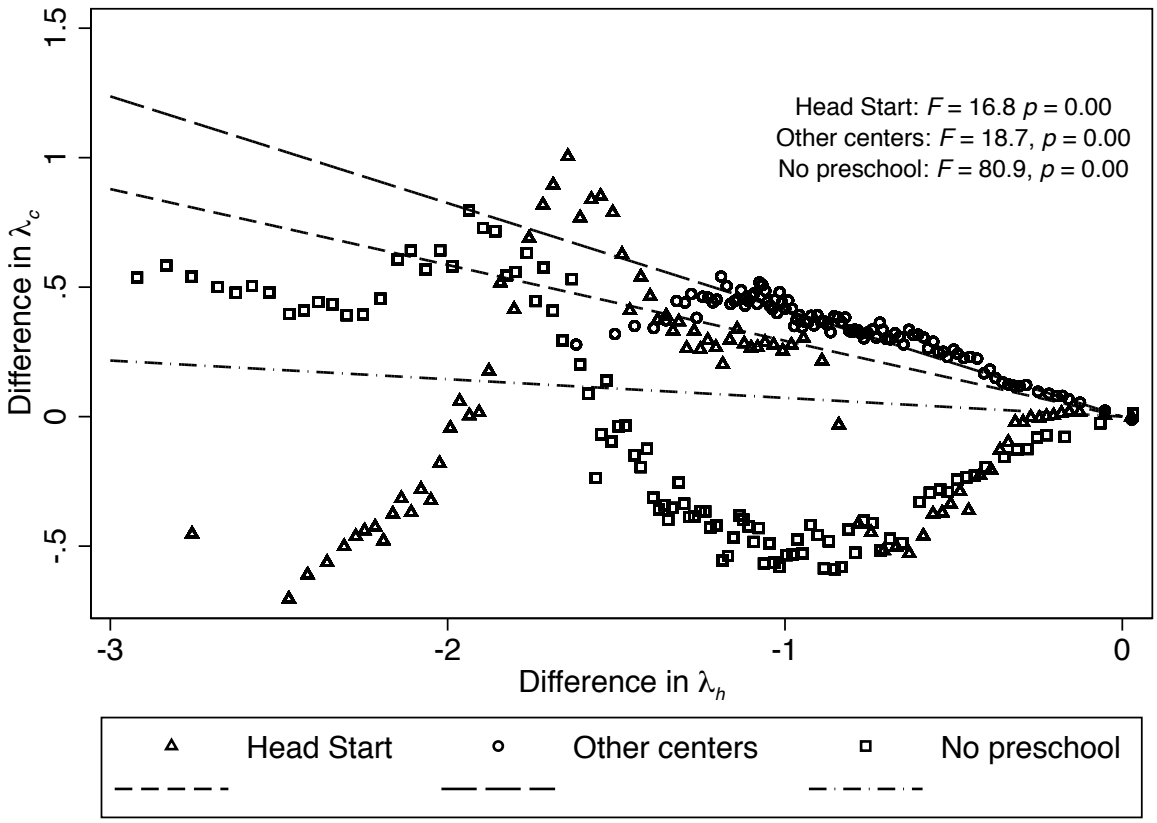


Figure A.II. Identification of the Selection Model

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

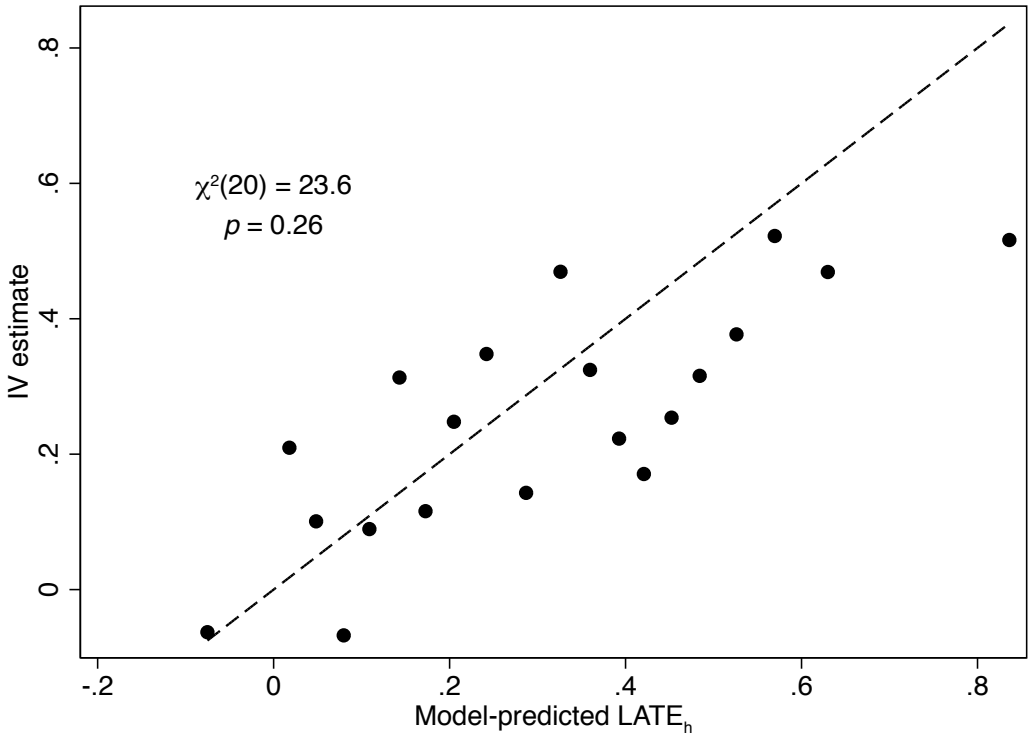


Figure A.III. Model-predicted $LATE_h$ vs. IV estimates

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

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

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

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

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

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

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

Table A.III. Effects on Maternal Labor Supply

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

| | | One endogenous variable | Two endogenous variables | |
|---|---|---|---|---|
| | | Head Start | Head Start | Other centers |
| Instruments | Estimator | (1) | (2) | (3) |
| Offer | 2SLS | 0.247 | - | - |
| (1 instrument) | | (0.031) | | |
| Offer x sites | 2SLS | 0.210 | 0.213 | 0.008 |
| (183 instruments) | | (0.026) | (0.039) | (0.095) |
| | First-stage *F* | 215.1 | 90.0 | 2.7 |
| | Overid. *p*-value | 0.002 | 0.002 | |
| | LIML | 0.218 | 0.029 | -0.581 |
| | | (0.027) | (0.139) | (0.432) |
| | Overid. *p*-value | 0.002 | 0.076 | |
| | JIVE | 0.217 | 0.109 | -0.329 |
| | | (0.026) | (0.110) | (0.332) |
| | Overid. *p*-value | 0.001 | 0.003 | |

Notes: This table reports two-stage least squares estimates of the effects of Head Start and other preschool centers in Spring 2003. The model in the first row instruments Head Start attendance with the Head Start offer. Models in the remaining rows instrument Head Start and other preschool attendance with interactions of the offer and indicators for experimental sites. Sites with fewer than 10 observations are grouped together within program areas as described in Appendix D. All models control for main effects of the interacting variables and baseline covariates. JIVE refers to the JIVE2 estimator defined in Angrist, Imbens and Krueger (1995), computed after first partialing out the exogenous covariates as described by Ackerberg and Devereux (2009). Overidentification tests for JIVE are based on Hansen's (1982) *J*-statistics for 2SLS and LIML. Overidentification tests for JIVE are based on the many instrument and heteroskedasticity-robust statistic derived by Chao et al. (2014). First stage *F*-statistics are Angrist/Pischke (2009) partial *F*'s. Standard errors are robust to heteroskedasticity.

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

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

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

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

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

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