# Selection into Identification in Fixed Effects Models, with Application to Head Start

Douglas L. Miller*
Cornell University
and NBER

Na'ama Shenhav†
Dartmouth College

Michel Z. Grosz‡
Abt Associates

June 6, 2019§

**Abstract**

Many papers use fixed effects (FE) to identify causal impacts of an intervention. In this paper we show that when the treatment status only varies within some groups, this design can induce non-random selection of groups into the identifying sample, which we term *selection into identification* (SI). We begin by illustrating SI in the context of several family fixed effects (FFE) applications with a binary treatment variable. We document that the FFE identifying sample differs from the overall sample along many dimensions, including having larger families. Further, when treatment effects are heterogeneous, the FFE estimate is biased relative to the average treatment effect (ATE). For the general FE model, we then develop a reweighting-on-observables estimator to recover the unbiased ATE from the FE estimate for policy-relevant populations. We apply these insights to examine the long-term effects of Head Start in the PSID and the CNLSY. Using our reweighting methods, we estimate that Head Start leads to a 3.1 percentage point (p.p.) increase (s.e. = 6.1 p.p.) in the likelihood of attending some college for white Head Start participants in the PSID. This ATE is 70% smaller than the traditional FFE estimate (12 p.p). We find qualitatively similar attenuation of the CNLSY estimates.

*Doug Miller, Policy Analysis and Management, Cornell University, Email: dlm336@cornell.edu
†Na'ama Shenhav, Department of Economics, Dartmouth College, E-mail: naama.shenhav@dartmouth.edu
‡Michel Grosz, Abt Associates, E-mail: michel_grosz@abtassoc.com
§We would like to thank Colin Cameron, Liz Cascio, Janet Currie, Hilary Hoynes, Pat Kline, Erzo F.P. Luttmer, Jordan Matsudaira, Zhuan Pei, Maya Rossin-Slater, Doug Staiger, Dmitry Taubinsky, Chris Walters, and participants at the AEA Meetings, Cornell, Dartmouth, CSWEP CEMENT Workshop, Mcgill University, NBER Labor/Children's Summer Institute, Northwestern, SEA Meetings, SOLE, Syracuse/Cornell Summer Workshop in Education and Social Policy, UC Merced, and the War on Poverty Conference at the University of Michigan. We are also grateful to Alex Magnuson, Wenran Li, and Wenrui Huang for excellent research assistance.

# 1 Introduction

Fixed Effects (FE) are frequently used to obtain identification of the causal impact of an attribute, intervention, or policy – the "treatment" of interest. This class of models has been used to identify the impact of academic peers (school-grade FE; Hoxby, 2000; Carrell and Hoekstra, 2010); criminal peers (facility-offense FE; Bayer, Hjalmarsson and Pozen, 2009); the local health care environment (individual FE; Finkelstein, Gentzkow and Williams, 2016); participation in means-tested programs (family FE; Currie and Thomas, 1995; Garces, Thomas and Currie, 2002; Deming, 2009; Rossin-Slater, 2013); neighborhood quality (family FE; Chetty and Hendren, 2018*a*); and minimum wage laws (county-pair-year FE; Dube, Lester and Reich, 2010), to give a few examples. Many of the estimates in these studies are naturally read as the average effect for a policy-relevant population (e.g. participants or those eligible for treatment). However, in contrast with other common estimators, there is not yet a comprehensive framework for considering the *external validity* of FE estimates.

In this paper, we show that FE can induce a special type of (non-random) selection in estimation, which we term *"selection into identification" (SI).* Broadly speaking, SI results from the fact that FE estimates are identified from FE groups (e.g. families, in the case of family FE) that have variation in treatment ("switchers"), which may exclude some groups.[1] In the contexts we examine, switchers are *(i)* a subset of the sample and *(ii)* systematically different than the overall population. This is a distinct problem from whether within-group comparisons are internally valid, which has been the typical subject of debate for FE estimators,[2] and which is not the focus of this paper. It is also different from the issue of conditional variance weighting of switcher treatment effects, which can also create external validity concerns (Gibbons, Suarez and Urbancic, 2018). We show that in the presence of heterogeneous treatment effects, SI causes FE to deviate from the ATE, and we develop reweighting-on-observables methods that can be used to recover the ATE for the overall population or for target populations (such as program participants). We apply these methods to revisit prior FE estimates of the long-run impact of Head Start.

We begin by presenting four facts that illustrate the empirical relevance of SI, in the context of a family fixed effects (FFE) model with a binary treatment. In particular, we examine patterns of within-family variation in participation Head Start, a federally-funded preschool program, using the Panel Study of Income Dynamics (PSID), as in Garces, Thomas and Currie (2002) (hereafter GTC).[3] First, relative to an estimation model without fixed effects, FFE uses substantially fewer identifying groups, more so than is commonly noted in work on this topic. Among the 5,355 children

[1] In the presence of control variables that vary within a group, then there may be variation among non-switchers "net of controls." We focus on cases where this phenomenon is small in magnitude, and formalize this extension in Section 5.1.

[2] See Bound and Solon (1999).

[3] Similar FFE models have been used to evaluate many other treatments; for public housing, see Andersson et al. (2016); for WIC, see Chorniy, Currie and Sonchak (2018); Currie and Rajani (2015); for health, see Almond, Chay and Lee (2005); Figlio et al. (2014); Abrevaya (2006); Black, Devereux and Salvanes (2007); Xie, Chou and Liu (2016), among others. We summarize the prevalence of this design in Section 2.

in the sample with siblings, only 1,098 children reside in switcher households. Second, the loss of sample variation is systematically related to observables. The likelihood of being a switcher – and thus being included in the FFE estimation – increases as the probability of treatment approaches 0.5, and with the number of units per group (children in a family). Third, since these factors vary across subgroups, SI does as well. The FFE identifying sample misses 93% of the sibling sample for white children, but only 62% of the sample for black children. Fourth, as a result, switchers are not representative of the overall sample along many dimensions. The most striking imbalance is along family size, but differences in income and parental education are also apparent.

Next, we show that under heterogeneous treatment effects, SI can meaningfully change the estimated treatment effect. The consequence of this is that the FFE estimate is no longer representative of the sample Average Treatment Effect (ATE), let alone the treatment effect for a policy-relevant population, such as program participants. This also implies that the difference between the OLS estimate and FE estimate can no longer be interpreted as solely reflecting OLS bias, even after accounting for conditional variance weighting among switchers. We show that this is a quantitatively more imporant source of bias in our applications than the bias from conditional variance weighting.[4] Because FE groups are less likely to be switchers when they are defined over a smaller groupings, the impact of SI may be stronger in those cases. Hence, in some settings standard FE methods may lead to a tradeoff between external and internal validity.

To address this, in Section 4 we take advantage of the insight that switching is a form of selection to develop a novel reweighting approach that can recover the ATE of policy-relevant "target" populations. Building on extrapolation methods designed to address non-representative experimental participants and IV compliers,[5] we show that the appropriate group-level weight for FE is proportional to the ratio of two propensity scores: (i) the propensity to be in the target population (e.g. program participants) and (ii) the propensity to be in the switcher population. Under the additional assumptions that these propensity scores can be estimated using observable covariates, and that unobservable determinants of switching are not correlated with treatment effects, we can then obtain the desired ATE.[6]

We demonstrate the performance of our reweighting using Monte Carlo simulations in a setting with naturally-occurring SI, which allows us to test the feasibility of our baseline modeling assumptions. We find that reweighting reduces or eliminates bias relative to FE in the presence of covariate-based treatment heterogeneity. In Section 5, we discuss several extensions of our basic setup, such as how the inclusion of covariates that vary within a group can create additional "residual switchers," and show how reweighting can be applied to a non-linear model.

Based on these findings we propose new standards for practice when presenting results using FE

[4]This is consistent with Gibbons, Suarez and Urbancic (2018), whose findings suggest that the bias from conditional variance weighting is less than 5% for 75% of estimates.

[5]See Angrist and Fernandez-Val (2013) for extrapolation from IV, and Stuart et al. (2011) and Andrews and Oster (2019) for extrapolation from experiments.

[6]In some settings, this assumption can be tested by comparing treatment effects across target and non-target populations within the switching sample, as we discuss in Section 4.

research designs: (i) clearly show the sample size when limited to switcher families and quantify the contribution of "residual switchers"; (ii) show the balance of covariates across switcher and non-switcher families (e.g. Table 2); (iii) reweight FFE estimates for a representative population (e.g. Table 6). Reweighted estimates can be presented either as an additional diagnostic tool or as an alternative measure of treatment effects. We are not the first to use the more rigorous reporting standards in (i) and (ii), but in our survey of the FFE literature the vast majority do not discuss either of these issues (e.g. one paper included (ii).)[7]

In Section 6, we apply these methods to quantify the importance of selection into identification for FFE estimates of the long-run impact of Head Start. Head Start has a budget of $8.6 billion dollars and annually enrolls roughly 60% of the number of 3 and 4 year old children in poverty, which makes it a quantitatively important intervention for this population (Carneiro and Ginja, 2014).[8] FFE have been used to identify the long term impacts of Head Start in many of the foundational studies of this program (Currie and Thomas, 1995; Deming, 2009, GTC), which find positive impacts on economic and non-cognitive outcomes of participants measured in adulthood. We provide new evidence of these effects, and also for the first time estimate the average long term effects for the Head-Start-eligible and Head-Start-participant populations.

Using data from the PSID and the Children of the National Longitudinal Study of Youth (CNLSY) (as in GTC and Deming (2009)), we newly document that, across multiple human capital measures, there are patterns consistent with greater returns to Head Start in larger families. This might result from the fact that parental time investment in children's human capital is spread more thinly in larger families, which in turn could lead to greater returns to alternative investments, such as Head Start.[9] Since these families are upweighted in FFE models, it is intuitive that the FFE estimate is likely to be upward-biased.

We illustrate the impact of reweighting first using the PSID and the largest sample of siblings – three times as large as the analysis in GTC – used to investigate this question. The FFE estimate in the PSID suggests that Head Start leads to a statistically significant 12 p.p. increase in attendance of some college. Using our reweighting methods, however, we find more modest and less-precisely-estimated benefits of the program.

Reweighting the estimates, we find that Head Start leads to a 3.1 percentage point (p.p.) increase in the likelihood of attending some college for Head Start participants (s.e. = 6.1 p.p.), and a 7.1 p.p. (se = 6.0 p.p.) increase for the Head Start eligible population. The ATE for Head Start participants estimate is 74% smaller than the FFE estimate, a difference which is significant

[7] Important exceptions include Finkelstein, Gentzkow and Williams (2016) and Wiswall (2013), who include a substantive discussion and examination of external validity concerns, as well as Currie and Rossin-Slater (2013). GTC report the number of identifying observations used to identify Head Start for the entire sample (not for subsamples), and Deming (2009) reports the aggregate number of identifying observations used to identify the pre-school, Head Start, and no-formal-care coefficients, but not for each coefficient.

[8] See Gibbs, Ludwig and Miller (2013) for an overview of the Head Start program.

[9] In Section 6, we show that this heterogeneity by family size is not explained by other covariates or by larger families having longer sibling cohort spans. Instead it appears that there is something important about family size per se.

at the 5 percent level. It is also 89% smaller than the estimated effects on college-going in GTC for this population; 22% to 89% smaller than unadjusted estimates for all participants from other FFE studies (Bauer and Schanzenbach, 2016; Deming, 2009); and 42% smaller than estimates from the county roll-out of Head Start (Bailey, Sun and Timpe, 2018), although the lower end of the confidence intervals for the latter estimates include our ATE.

Reweighting similarly attenuates the FFE estimate of the impact of Head Start in the CNLSY (Deming, 2009). While the FFE estimate suggests that Head Start leads to an 8.5 p.p.increase in high school completion, the reweighted estimate for Head Start participants is 40% smaller and significant only at the 10% level. The FFE and reweighted estimates are statistically different at the 10% level. Reweighting also attenuates the previously-estimated impact of Head Start on idleness and having a learning disability and, to a lesser degree, the impact on poor health, relative to the FFE estimates.

We present our results primarily in the context of FFE and Head Start, but they apply to any panel fixed effects model, with special relevance for those with short panels and "lumpy" treatment variables (e.g. binary treatments). For instance, we use data from Collins and Wanamaker (2014) to demonstrate similar patterns of selection into identification in the estimation of returns to migration with FFE.

In Section 7, we discuss three additional potential applications of our methods to FE estimation of peer effects (school-grade FE; Carrell, Hoekstra and Kuka, 2018; Carrell and Hoekstra, 2010), the minimum wage elasticity (county-pair-year FE; Dube, Lester and Reich, 2010), and responses to environmental shocks (district FE; Shah and Steinberg, 2017). We identify features within each of these settings that make the estimates potentially subject to selection into identification. As a result, we recommend careful investigation of these issues in future research using FE strategies.

Two of the core contributions of this paper are first to provide guidelines that can be used to characterize the likelihood of being a non-switcher (based on the probability of treatment or the number of units in a group); and second to show the importance of heterogeneity in treatment effects across switching and non-switching groups. While it is well-known that the FE estimator is only identified from switchers, we document in a review of the literature that the number of switchers and their characteristics is not commonly discussed in applied work. We show in our applications that non-switching is common, and that switching groups are not randomly distributed in the population. This has a meaningful impact on the external validity of estimates.

The prevalence of non-switching also stands in contrast to a commonly-held assumption of positive within-group variance of treatment used for theoretical findings; such as in the translation from FE to IV (Loken, Mogstad and Wiswall, 2012), and in the reweighting of OLS (Angrist, 1998), IV (Angrist and Fernandez-Val, 2013; Aronow and Carnegie, 2013) and FE (Gibbons, Suarez and Urbancic, 2018) for external validity.

Third, we provide a reweighting estimator that allows for the recovery of ATE for policy-relevant populations. This is different from strategies that employ reweighting for internal validity,

such as traditional propensity score estimation methods, and from recent works on the validity of difference-in-difference strategies, where the empirical specification ensures that SI is unlikely to be a concern.[10]

Our reweighting solution is broadly related to a growing literature that identify and correct for the discrepancy between "what you want" and "what you get" from common estimators, such as Lochner and Moretti (2015), who reweight OLS with IV weights for greater comparability; Sloczynski (2018), who reweights OLS to obtain the ATE; and Stuart et al. (2011) and Andrews and Oster (2019) who propose reweighting experiments to account for selection into participation.

Closest to the current work are reweighting strategies for quasi-experimental estimates, including Angrist and Fernandez-Val (2013), who reweight IV using discrete covariates, and Gibbons, Suarez and Urbancic (2018), who reweight FE using inverse-conditional-variance weights to obtain the switcher ATE. Unlike Angrist and Fernandez-Val (2013), we focus on the external validity of FE and reweight using a propensity score, which allows for greater flexibility in conditioning variables. Further, our reweighting method relaxes the assumption of positive-conditional-variance in Gibbons, Suarez and Urbancic (2018), and provides a means for extrapolating from switchers to a policy-relevant target population. This should be informative for treatment estimates, since switchers are not typically a population of interest.

Finally, we contribute to a growing body of work investigating the long term effects of Head Start using quasi-experimental methods (Ludwig and Miller, 2007; Carneiro and Ginja, 2014; Thompson, 2017; Bauer and Schanzenbach, 2016; Johnson and Jackson, 2017; Bailey, Sun and Timpe, 2018; Pages et al., 2019; Barr and Gibbs, 2018, in addition to the FFE papers above). These studies typically present LATE or ITT estimates, and find improvements in childhood health, reductions in adolescent behavioral problems and obesity, and increases in adult educational attainment and earnings.[11] Relative to most of these studies, we evaluate the effect of Head Start on longer-run outcomes; show that these effects vary significantly by family size; and also adjust estimates using covariate re-weighting to get closer to the ATE for Head Start participants. We show that incorporating this adjustment lowers the estimated long term effect of Head Start.

## 2 A Survey of FFE Applications

Since our application focuses on a FFE model, we focus on applications of this particular method in the literature. This focus will lead us to undercount the prevalence of FE more broadly, but provides an unambiguous example of a short-panel setting which is susceptible to SI concerns. We surveyed publications from January 2000 to May 2017 in 11 leading journals that publish applied microeconomics articles. We include all studies that use family fixed effects as a primary

[10] See, e.g. Goodman-Bacon (2018); Borusyak and Jaravel (2017); Callaway and Sant'Anna (2018). We discuss SI in two-way FE designs in Section 5.1.

[11] One exception to this is Pages et al. (2019), who suggest that the effect of Head Start may be negative for recent cohorts, although the identifying sample is not discussed.

or secondary strategy.[12]

Our literature review yields 55 papers published from 2002 to 2017. We provide descriptive statistics of these articles in Table 1, including statistics by journal. Overall, these articles account for less than 1 percent of the papers published in our sample of journals, but this varies from 0 to 3 percent of each journal. The first panel tabulates the frequency of binary treatments and binary outcomes across the sample of papers, the focus of our methodological insights. Nearly two-thirds (35) of the papers have a binary treatment of interest and 23 have a binary outcome. The second and third panels show the varied topics that appear in the sample, spanning health, public, education, and labor fields.

The final panel of the table summarizes the distribution of sample sizes used with FFE. The samples are frequently not limited to families with variation in the treatment variable; therefore, the sample size in the table is an upper bound on the number of observations used for identification. The median number of sibling observations is 6,315, or roughly 85% of the sample in our analysis. We note that there is a high variance in sample size across samples, indicating that there is not a threshold for FFE analyses. The bottom 25% of papers have fewer than 1,200 observations, while the top 25% have over 160,000 sibling observations.

Appendix Figure B.1 illustrates the popularity of this estimation strategy over time. It shows a steady stream of FFE papers over the past 15 years; and that these papers have an impact on the literature, with a mean 233 citations per article (Google Scholar citations as of May 2019). Moreover, since the survey was completed, additional FFE studies have been published (see e.g. Chetty and Hendren (2018*a*,*b*)).

## 3 Fixed Effects and Selection into Identification

We employ the FE research design to address the concern that Head Start treatment may be correlated with some fixed characteristics of a family that also determine outcomes. For example, the decision to participate in Head Start of siblings is influenced by low parental income — a requirement for eligibility — and availability of an alternative source of care, which may independently influence long-term outcomes. As a result, in our setting – as well as in many other settings – the cross-sectional estimate of the effect of treatment is likely to be biased.

To formalize our setting, let $D_i \in \{0, 1\}$ indicate whether an individual $i$ participates in treatment (e.g. Head Start) and $g(i)$ be the relevant group (e.g. family) for $i$ of the set of groups $G$ in the sample, and let potential outcomes in the untreated and treated states be $Y_i(0)$, $Y_i(1)$, respectively.

[12] We surveyed: AEJ: Applied Economics, AEJ: Economic Policy, AER, AER P&P, Journal of Health Economics, Journal of Human Resources, Journal of Labor Economics, Journal of Political Economy, Journal of Public Economics, QJE, Review of Economics and Statistics. To identify these articles, we used the search terms "family," "within-family," "sibling," "twin," "mother," "father," "brother," "sister," "fixed effect," "fixed-effect," and "birthweight" using queries on journal websites. We then searched within articles to see whether FFE was used in the analysis. Finally, we added some additional papers to the list that we are aware of and did not satisfy these search terms. The resulting list is fairly comprehensive, but still likely to be a slight undercount of FFE articles in these journals.

We observe for each $i$ one outcome, $Y_i = Y_i(D_i)$, treatment, $D_i$, and group membership, $g(i)$. For brevity, we will frequently write this simply as $g$. We refer to groups for whom $Var(D_i|i \in g(i)) > 0$ as "switchers," and denote switching status with a binary variable $S_g = 1$, and the set of switchers as $G_S \subseteq G$.

We assume that although treatment may be correlated with group characteristics, e.g. mean family income, it is randomly assigned within groups:

**Assumption 1 (Group ID Conditional Independence):**

$$Y_i(0), Y_i(1) \perp\!\!\!\perp D_i|g(i) = g \tag{1}$$

Assumption 1 encompasses the standard FE specification assumpton in linear models. It rules out Roy (1951)-type selection into treatment within groups, in which treatment is correlated with treatment effects, $Y_i(1) - Y_i(0)$).[13] In the context of Head Start, treatment has been shown to be uncorrelated with most observable characteristics of children (Deming, 2009, GTC, 2002), suggesting the assumption is reasonable.

Under this assumption, estimated treament effects $\hat{\delta_g}$ are an unbiased estimate of group-level treatment effects, $\delta_g \equiv \mathbb{E}[Y_i(1) - Y_i(0)|g(i) = g]$. The FE estimate averages $\hat{\delta}_g$ for the $g \in G_S$, using weights that are proportional to the within-group variance of $D_i$ and the number of observations in $g$ (Angrist, 1998; Angrist and Pischke, 2009, eqn. 3.3.7), as follows:

$$\delta_{FE} = \sum_{g \in G_S} \delta_{g,FE} \cdot \omega_{g,FE} \tag{2}$$

where

$$\omega_{g,FE} = \frac{Var(D_i|g(i) = g) \cdot Pr(g(i) = g|S_g = 1)}{\sum_{g \in G_S} Var(D_i|g(i) = g) \cdot Pr(g(i) = g|S_g = 1)}$$

We examine two methodological issues that arise from the FE research design: *(i)* reduction in identifying variation moving from $G$ to $G_S$; and *(ii)* a change in the composition of the identifying sample. Issue (i) is well understood in principle, but the degree to which $G_S$ is smaller than $G$ is often underappreciated, not reported in empirical practice, and implicitly assumed to be negligible in theoretical results. Issue (ii) is more novel, and should cause researchers to update the interpretation of the population for which these estimates are relevant.

### 3.1 Empirical Relevance

To illustrate ideas, we provide an empirical example - for more detail, see Section 6. The sample consists of 2986 white children born in the years 1954-1987. The regression of interest

[13] Some recent FE strategies explore relaxation of this assumption. For example, in a two-period person-level FE design, Lemieux (1998) estimates union wage returns to both observed and unobserved skills. This approach is extended (with application to farmer adoption of HYV seeds) in Suri (2011) and Verdier and Castro (2019).

estimates the effect of ever having attended Head Start on a dummy for ever having attended college. The coefficient on Head Start in a cross-section regression is 0.049 (s.e. = 0.044). When mother fixed effects are added, the coefficient becomes 0.120 (s.e. = 0.053). This result indicates that the impact of Head Start participation on college attendance is meaningful in magnitude, and statistically significantly different from zero.

We illustrate the identifying variation for the FFE regression of some college on Head Start attendance in Panel (a) of Figure 1, which shows a scatterplot of the deviation in Head Start attendance for each individual $i$ from the mean attendance in his or her family, $g(i)$, $HeadStart_i$-$\overline{HeadStart_{g(i)}}$, against the within-family deviation in attainment of some college for the sample, $AnyCollege_i$-$\overline{AnyCollege_{g(i)}}$.[14] Strikingly, the largest mass of observations is at (0,0): the majority of families have no variation in Head Start participation and no variation in the college attendance of their children. Individuals in families with no variation in Head Start account for 96% of the sample – removing these leaves us with 213 individuals in switching families.

This reduction in identifying observations could result in a selected sample if switching is correlated with family characteristics. To gain intuition about which variables might determine switching we build a simple model of the Head Start (HS) participation decision within families. If the probability of attending Head Start is a constant, $\pi$, and independent across siblings in a family, then the probability of switching, $P(S_g = 1)$ is simply a function of $\pi$ and family size, $n_g$:

$$Pr(S_g = 1) = 1 - (1 - \pi)^{n_g} - \pi^{n_g}$$

According to this formula, the probability of switching has an inverse-U-shaped relationship with $\pi$, peaking at $\pi = 0.5$. Further, for a given level of $\pi$, the likelihood of being in a switching family is increasing with family size. We illustrate these features in Appendix Figure B.2.

The markers in Figure 2 show the actual probability of attending Head Start and of being in a switching family for each family size by black/white race and by whether the mom has some college or not. As in the stylized model, the likelihood of switching is increasing with family size for each of these subgroups.[15] This could reflect the fact that over time, across children, parents are more likely to be exposed to the program, or are more likely to experience a change in family income, which alters eligibility for the program.

We also observe that switching increases with $\pi$, following the inverse-U. The probability of Head Start attendance among black families and families with low-educated moms is much higher and closer to 0.5, compared to white families and families with high-educated moms; and the switching probability is correspondingly larger for black and low-educated families. As a result, the

[14] The size of each symbol is weighted by the number of individuals. A value of 0.5 along the horizontal axis, for example, means that a person went to Head Start in a family where half the children attended Head Start. Values other than 0.5 and -0.5 indicate that the share of children that attended Head Start was different than 0.5; e.g. a value of -0.75 means that a person did not go to Head Start in a family where three quarters of the children did.

[15] Appendix Table B.1 shows that this pattern is driven by a much larger incidence of no Head Start participation among smaller families. For example, 78% of 2-child families have no Head Start participants, compared with 48% of families with 5 or more children.

sample used for FFE identification is comprised of 7% of the sibling sample for whites, and 38% of the sibling sample for blacks. Note that while we are focusing on race and maternal education, this notion can be generalized to any other family characteristic, such as SES, that determine $\pi$.

This pattern is not unique to the PSID or to Head Start. Panels (b) and (c) of Figure 2 show this relationship using data from two other FFE papers, Collins and Wanamaker (2014) and Deming (2009). In both papers, the treatment variable of interest is binary; migration to the North and Head Start participation, respectively. In each of these samples, the probability of being a switcher is increasing in family size.

### 3.1.1 Selection into Identification Driven by Many Variables

Since SI is likely to affect the balance of characteristics other than family size, we now examine a large number of observable characteristics of switcher families and non-switcher families. Panel A of Table 2 indicates that in addition to having a larger family size, children in switcher families tend to have parents with significantly less education than children in non-switcher families (column 3). These differences in parental education are significant even in a regression framework where we control for differences in family size and the other covariates in the table, though only at the 10 percent level (columns 4 and 5). Family income during preschool of children in switcher families is significantly lower than non-switcher families overall (some of which may have incomes too high to ever qualify for Head Start).[16] These patterns are consistent with switching increasing with the probability of Head Start participation.

Next, we examine a one-dimensional summary of how much overlap there is in the characteristics of switchers and non-switchers. We do so by constructing a propensity-score-type summary measure, $\frac{Pr(S_g|X_{ig})}{Pr(HS|X_{ig})}$, which gives a measure of how aligned the characteristics ($X_{ig}$) of switchers are with the characteristics of Head Start participants, the population of interest. An average value of 1 implies perfect alignment, while a higher value implies that the characteristics of switchers are over-represented relative to the characteristics of Head Start participants. We estimate the elements of this ratio using a multinomial logit.

Panel B of Table 2 shows that this measure is between 1.9 and 2.9 for the switchers sample, which is 0.3 SD larger than for non-switchers.This indicates that the observables of switchers are not aligned with our population of interest, and that this misalignment is worse for switchers than non-switchers.[17]

[16] If we limit ourselves to families with Head Start participants, we obtain qualitatively similar results, but the differences are somewhat smaller and sometimes less precisely estimated.

[17] As a benchmark, Stuart et al. (2011) suggest that a 0.1 to 0.25 SD difference in propensity scores between the experimental and non-experimental population may be too large to rely on extrapolation without further adjustments.

### 3.2 Consequences for Estimation: Effective Number of Identifying Observations

A convenient way to summarize the amount of variation used in FE is by the number of individuals in switching families. However, since not all switchers provide the same amount of identifying variation, this can be a misleading measure. For example, a 4-sibling family with 1 treated and 3 untreated individuals has an $\omega_{g,FE}$ that is 25% smaller than the $\omega_{g,FE}$ of a family with 2 treated and 2 untreated ($0.25 \cdot 0.75 = 0.1875 < 0.25 = 0.5 \cdot 0.5$).

We develop a formula for the "effective number of observations," which captures this idea by (i) quantifying the total amount of identifying variation and (ii) converting this into standardized units (person-equivalents).

$$N_{eff} = \frac{\sum_{g \in G_S} Var(D_i|g(i) = g) \cdot [n_g - 1]}{Var(D_{i,reference})} \tag{3}$$

The numerator quantifies the "total amount of variation" identifying $\delta_{FE}$. Different from the FE formula, family size is adjusted for the fact that group-level fixed effects remove one degree of information from each family, $[n_g - 1]$. The denominator provides a translation from "total variation" to "person-equivalents" of variation by normalizing by the variation contributed by an individual observation in a fixed, researcher-determined group, $Var(D_{i,reference})$.

In our application, we report effective observations using as reference (i) the variation in a cross section regression after controlling for reasonable g-level covariates, $Var(D_{i,reference})$ $=Var\left(D_i|W_g\right)$; and (ii) the variation from individuals in groups in two-child families.[18]

### 3.3 Consequences for Estimation: Bias

Under homogeneous treatment effects ($\delta_g = \delta$), SI has no effect on expected bias in estimation of Equation 2, and the FE estimate trivially is unbiased for the ATE for the sample and the population. There is only a loss of precision that accompanies the overall reduction in sample size.

The more interesting case is when treatment effects are heterogeneous. In that case, SI will lead the FE estimate to provide a biased estimate of the ATE, even if one corrects for the conditional variance weighting of FE among switchers. To be concrete, let $Z$ be a discrete covariate that varies at the group level, such as family size, that determines the magnitude of the effect of treatment. We allow for a different treatment effect for eachvalue of $Z$: $\delta_g = f(z_g) = \delta_z$, and define $\mathbb{Z}$ as the set of values of $z_g$ present in the samples of siblings and switchers. The treatment effect estimated without FE using a sample of groups with $n_g \geq 2$, e.g. siblings, is:

$$\delta_{OLS} = \sum_{z \in \mathbb{Z}} \delta_{z,OLS} \cdot \omega_{z,OLS} \tag{4}$$

[18] $Var(D_i|g(i)$ is calculated using the population formula for variance, $Var(D_i|g) = \frac{1}{n_g} \sum_{i \in g} \left(D_i - \overline{D_g}\right)^2$, rather than the sample formula (which would divide by $n_g - 1$).

where

$$\omega_{z,OLS} = \frac{(Var(D_i|n_g \geq 2, z_g = z) \cdot Pr(z_g = z|n_g \geq 2)}{\sum_{z' \in \mathbb{Z}}(Var(D_i|n_g \geq 2, z_g = z') \cdot Pr(z_g = z'|n_g \geq 2)}$$

and $\delta_{z,OLS}$ is the OLS estimate without FE of the treatment effect for groups with $z_g = z$, and $Var(D_i|n_g \geq 2, z_g = z)$ is the conditional variance of treatment among the sample with $n_g \geq 2$ and $z_g = z$.

The FE estimator for the same sample is:

$$\delta_{FE} = \sum_{z \in \mathbb{Z}} \delta_{z,FE} \cdot \omega_{z,FE} \tag{5}$$

where

$$\omega_{z,FE} = \frac{(Var(D_i|FE, z_g = z) \cdot Pr(z_g = z|S_g = 1)}{\sum_{z' \in \mathbb{Z}}(Var(D_i|FE, z_g = z') \cdot Pr(z_g = z'|S_g = 1)}$$

and $\delta_{z,FE}$ is the FE estimate of the treatment effect for groups with $z_g = z$, $Var(D_i|FE, z_g = z)$ is the conditional variance of treatment among the sample for groups with $z_g = z$, net of family fixed effects.

Moving from OLS to FE, the $\delta$'s change and also the $\omega$'s change. The change in the $\delta$'s is how we usually interpret the move from OLS to FE: the change is from "between" (bad) variation to "within" (good) variation. But the full change also incorporates the different weightings of different values of $z_g$. If the OLS sample and the FE sample overlap in the covariates, we can decompose the difference between OLS and FE to identify how much is caused by the change in weights, $\omega_z$, and how much is driven by the change in identification, $\delta_z$, as:

$$\begin{aligned}\delta_{FE} - \delta_{OLS} = &\sum_{z \in \mathbb{Z}} \underbrace{[\omega_{z,FE} - \omega_{z,OLS}] \cdot (\alpha \cdot \delta_{z,FE} + (1-\alpha) \cdot \delta_{z,OLS})}_{\text{Impact of } \Delta \text{ weighting}} \\ &+ \sum_{z \in \mathbb{Z}} \underbrace{(\delta_{z,FE} - \delta_{z,OLS}) \cdot (\alpha \cdot \omega_{z,OLS} + (1-\alpha) \cdot \omega_{z,FE})}_{\text{OLS Bias}}\end{aligned} \tag{6}$$

With $\alpha \in [0, 1]$ a researcher-determined weight. The impact of SI is captured in the first summation of Equation 6, which is a function of the disparity in regression weights $\omega_z$ between OLS and FE, multiplied by an $\alpha$-weighted average of the $\delta_{z,OLS}$ and $\delta_{z,FE}$. Setting $\alpha = 0$ in this term uses cross-section coefficients to assess the importance of changing the regression weights from OLS to FE. Setting $\alpha = 1$ uses the FE coefficients to assess this. If there is important heterogeneity among both $w_z$ and $\delta_z$, these two extremes can provide useful benchmarks to compare against the OLS and FE estimates, as we do in Section 4.3.[19]

[19] This decomposition is similar in form to Equation 13 in Loken, Mogstad and Wiswall (2012), which uses $\alpha = 1/2$. However, we sum over a group-level covariate that is distinct from the treatment of interest, while Loken, Mogstad

Since SI impacts the probability of each family size appearing in FE and OLS and possibly the conditional variance as well, existing methods to reweight FE estimates (Gibbons, Suarez and Urbancic, 2018) can at best recover the ATE for switchers. Since switchers are typically not a population of interest, this raises concerns for the external validity of the FE estimator.

**Illustration of Consequences: Greater Returns to Head Start in Larger Families**

We use data from our empirical example to illustrate the change in the components of $\omega_z$ across OLS and FE. Panel A of Table 3 shows that the proportion of 5+-child families in the switching sample is roughly twice the proportion in the overall sample, while the share of 3 and 4-child families is roughly similar. The variance in Head Start, shown in Panel B, is higher, roughly double, in the switching sample relative to the sibling sample, however this is relatively similar across family sizes. This suggests that the change in the conditional variance across OLS and FE plays a minor role in our setting.[20] We then calculate $\omega_{z,OLS}$ and $\omega_{z,FE}$. Going from the sibling sample to the switchers sample, $\omega_{2-child}$ declines by over 25% and $\omega_{3-child}$ declines by 15%. Conversely, $\omega_{5-child}$ nearly doubles from 0.134 to 0.243, and $\omega_{4-child}$ families increases by over 25%.

The effect of Head Start also varies by family size in our applications. The first two columns of Panel A of Table 4 shows the estimated effects of Head Start on the likelihood of completing some college by the number of children in a family for our illustrative sample. We show the results with and without family fixed effects. In both specifications, the effect of Head Start is significantly higher among white children in families with 5 or more children and, once fixed effects are added, the effect of Head Start is monotonically increasing with the number of children in a family.

One possible explanation for this heterogeneity is that children with higher initial endowments receive greater parental investments in larger families, and also benefit more from Head Start (Aizer and Cunha, 2012). Another possibility is that Head Start substitutes for parental time, which is more scarce in larger families. Another interpretation is that this heterogeneity reflects the fact that other covariates correlated with family size, such as income, mediate the impacts of Head Start. This final explanation seems less likely, as we find that the heterogeneity in family size survives the inclusion of other interactions, as we discuss in Section 6.

The bottom of Panel A shows the number of Head Start switcher observations and effective observations in terms of cross-sectional and two-sibling switcher individuals.[21] It shows that a total of 213 individuals are used to identify these coefficients, less than one tenth of the total sample, and that the variation is equivalent to 236 individuals in 2-person switching families. Hence, by including families with three or more children, on average, each observation is providing more

and Wiswall (2012) sum over values of an individual covariate (that varies within families), which is also the treatment of interest.

[20] We provide additional evidence that "undoing" the conditional variance weighting makes little difference in this application in Section 6.

[21] For effective cross-sectional individuals, the denominator of Equation 3 is the variance of Head Start, residualized by the family mean of the covariates in the analysis For the effective number of two-person switcher individuals, the denominator is $\left[V\left(D_i|g\right)\cdot\left(n_g-1\right)\right]/n_g = \left[0.5^2\cdot\left(2-1\right)\right]/2 = 0.125$.

variation than in a similar-sized sample of 2-child families. Further, the variation is equivalent to 731 individuals in a cross-sectional regression. This is because there is relatively little variation in Head Start in the full sample.

Like with SI, the larger Head Start effects we document for big families is not specific to the PSID. Columns (3) to (5) of Table 4 show the CNLSY FFE estimated effects of Head Start by family size for idleness, having a learning disability, and being in poor health.[22] For each of these outcomes, the impact of Head Start for 5+ child families is at least twice as large as the impact for 2 or 3 child families. For high school graduation, we also see a large impact for 4-child families, roughly double the impact for 2 and 3 child families. This implies that we should expect an increase in the coefficient going from OLS to fixed effects due to the *change in weighting* across the identifying samples, even without a change in the source of identification.

The number of switchers in the CNLSY sample is 581, less than half of the total number of observations. As in the PSID, the variation in this sample is equivalent to a larger sample of 2-person families (648 individuals.) However, the corresponding cross-sectional observations is smaller (438.7). These two examples illustrate that there are multiple forces driving the effective number of observations calculation: lost information from the group FEs drives down variation; but moving toward larger conditional variance of treatment increases variation. In the PSID example the second effect dominates; in the CNLSY case the first effect dominates.

# 4 Extrapolating from Identifying to Target Population

The difference between OLS and FE in implicit weighting of heterogeneous treatment effects leads us to consider translating the FE estimates into an ATE for a (researcher-determined) population of interest. We propose a method to flexibly obtain the ATE for such populations of interest, which we refer to as "target" populations, and denote by an indicator $T_g$. Commonly, the target population in applied work is the ATE for a nationally representative sample, which may be a reasonable starting place for most researchers. For some treatments, like means-tested programs, one might be interested in the ATE for eligible families, or families with a participating member.

## 4.1 Assumptions and Proposition

The methods rely on four key assumptions. which are variants of those used for extrapolation from IV (Angrist and Fernandez-Val (2013); Aronow and Carnegie (2013)). First, we assume that Group ID conditional independence (Assumption 1, Equation 1) holds.

**Assumption 2 (Conditional Fixed Effect Ignorability (CFEI)):**

[22] We focus on these outcomes because individuals that attended Head Start were found to fair significantly better on each of these outcomes in Deming (2009).

$$\mathbb{E}[Y_i(1) - Y_i(0)|S_g, P_x, Q_x] = \mathbb{E}[Y_i(1) - Y_i(0)|P_x, Q_x] \tag{7}$$

$$\mathbb{E}[Y_i(1) - Y_i(0)|T_g, P_x, Q_x] = \mathbb{E}[Y_i(1) - Y_i(0)|P_x, Q_x] \tag{8}$$

Second, we assume that conditional on observables, the true treatment effect is independent of a group's switching or target status, which we refer to as conditional fixed effect ignorability (CFEI).[23] We use two propensity scores constructed from the vector of group characteristics, $\mathbf{X_g}$, as the conditioning variables: $P_x := Pr\left[S_g = 1|\mathbf{X_g} = \mathbf{x}\right]$ is the propensity to be a switching group, and $Q_x := [T_g = 1|\mathbf{X_g} = \mathbf{x}]$ is the propensity to be in the (researcher-determined) target group. CFEI eliminates, for example, a second type of Roy (1951)-type selection, whereby switchers have an unobserved quality that increases the effectiveness of treatment.

In the Head Start application, the key determinants of $Pr[S_g = 1]$ are family size and the underlying probability of Head Start participation. Family size is observable, and observable covariates, such as family income, can take us a long way in predicting program participation. Likewise, the family-level determinants of $Pr[T_g = 1]$ for a target such as Head Start participants will be largely tied to observable eligibility requirements for the program, such as income and household size, which together determine the income-to-poverty ratio.

**Assumption 3 (Correct Propensity Score Specification):**

$$Pr(S_g = 1|\mathbf{X_g}) = F\left(\theta_\mathbf{g}; \mathbf{X_g}\right) \tag{9}$$

$$Pr(T_g = 1|\mathbf{X_g}) = G\left(\chi_\mathbf{g}; \mathbf{X_g}\right) \tag{10}$$

Third, we assume that the propensity scores that we estimate have the correct functional form, with $F\left(\cdot\right)$ and $G\left(\cdot\right)$ known, and $\theta_\mathbf{g}$ and $\chi_\mathbf{g}$ parameters to be estimated. In our application, we model $F\left(\cdot\right)$ and $G\left(\cdot\right)$ as a multinomial logit.

**Assumption 4 (Common Support):**

$$\text{If } Q_x > 0, \text{ then } P_x > 0, \forall \mathbf{x} \tag{11}$$

Fourth, we require positive probability of being a switcher for each value of $\mathbf{x}$ that is represented in the target group. Satisfying this assumption can place limits on the allowed target group. For example, covariates that perfectly predict "never switchers" such as singleton groups are ruled out of consideration as targets by this assumption. Extrapolating to these groups requires stronger functional form assumptions about the relationship between the treatment effect and the $X$'s. We are addressing this in an extension in progress.

[23] We considered using CoFEfe as the acronym for this assumption. This would provide a novel candidate interpretation of US President Donald Trump's enigmatic tweet of May 31, 2017.

**Proposition 1.** *Define the re-weighted FE estimator for target population $t$ as*

$$\widehat{\delta^t} := \frac{1}{\sum_i \mathbf{1}(S_{g(i)} = 1)} \sum_{i|S_{g(i)}=1} \widehat{w^t_{g(i)}} \cdot \hat{\delta}_{g,FE}, \tag{12}$$

*with $\widehat{w^t_{g(i)}}$ our estimate of $w^t_{g(i)}$,*

$$w^t_{g(i)} := \frac{Q_x \cdot Pr\left[T_g = 1\right]}{P_x \cdot Pr\left[S_g = 1\right]} \tag{13}$$

*Under Assumptions 1 through 4, $\widehat{\delta^t}$ is consistent for the ATE of the target population, $\mathbb{E}\left[Y(1) - Y(0)|T_g = 1\right]$.*

The proof is in Appendix A. Intuitively, the weights are increasing in $Q_x$ and decreasing in $P_x$, such that we upweight observations that are more similar to the target, and downweight observations that are overrepresented in the switching population. In other words, we weight the treatment estimate for each switcher group $g$ proportionately to match the share of the target population with observable characteristics matching $g$ (i.e the same $Q_x$ as $g$), which gives the ATE under the assumptions above.[24]

### Testable Implication

CFEI requires that treatment effects should be balanced across $T_g$, conditional on $P_x$ and $Q_x$. This is potentially testable if some switchers are not in the target population – for instance, if the target population is families that participate in a safety net program, groups that live in rural areas, or firms that are in a particular industry. We implement this in Section 6.2.1. This test can not be used, however, if the target is "everyone," "multi-unit groups," or otherwise contains the set of switchers $G_S$.

## 4.2 Reweighting Methodology

When the sample is a subset of the target population, $Q_x = 1$, and $P_x$ can be estimated by a logit or probit model. Otherwise, an observation might be in one or both (or neither) of the target population and the switching sample. We use a multinomial logit model to estimate the probability of each of the four possible combinations. $Q_x$ is then constructed as the sum of the predicted $Pr(T_g = 1, S_g = 0)$ and the predicted $Pr(T_g = 1, S_g = 1)$ for each unit; and $P_x$ is constructed as the sum of the predicted $Pr(T_g = 1, S_g = 1)$ and the predicted $Pr(T_g = 0, S_g = 1)$ for each unit.[25]

With these weights in place, the ATE for the target population can be estimated in one of two ways. The first is a two-step "post-regression weighting" of $\hat{\delta}_g$, where $\hat{\delta}_g$ is estimated from a regression of the outcome on interactions between $D_i$ and group-specific dummies. Then aggregate

[24] See Appendix A for a simple derivation of this equivalence.

[25] This can also be multiplied by survey weights, as we do in our PSID example.

$\hat{w}_g^t$ to the group-level and perform a normalization to obtain the final estimation weights, $\hat{s}_g^t = \frac{\hat{w}_g^t \cdot n_g}{\sum_{g \in G_s} \hat{w}_g^t \cdot n_g} = \frac{\frac{Q_x}{P_x} \cdot n_g}{\sum_{g \in G_S} \frac{Q_x}{P_x} \cdot n_g}$. The 2-step ATE combines these using:

$$\widehat{\delta_{2step}^t} = \sum_{g \in G_S} \hat{s}_g^t \cdot \hat{\delta}_g \tag{14}$$

Under the standard cluster-robust assumption that model errors are independent across groups, $\widehat{\delta_{2step}^t}$ is a weighted average of independent variables, and we can obtain a cluster-robust variance estimate as:

$$\widehat{Var}(\widehat{\delta_{2step}^t}) = \sum_{g \in G_S} (\hat{s}_g^t)^2 \cdot \left(\hat{\delta}_g - \widehat{\delta_{2step}^t}\right)^2 \tag{15}$$

A second approach is to obtain the ATE in a single step using "in-regression weights." For this, we need to adjust for the fact that the FE estimator uses weights $\omega_{FE}$ rather than population shares. We address this by incorporating inverse conditional variance weights, as $v_g = (Var(D_i|g(i) = g)^{-1}$ (Gibbons, Suarez and Urbancic, 2018).[26] Then, the ATE can be estimated by $\widehat{\delta_{1step}^t}$ from a one-step regression using $\widehat{w_{g(i)}^t} \cdot v_g$ as regression weights, and computation of cluster-robust standard errors is straightforward.[27]

### 4.3 Special Case: Univariate Heterogeneity

If the source of heterogeneity in estimates is a single, discrete covariate, we can obtain further insight from performing the decomposition captured in Equation 6. Taking the OLS family-size-specific coefficients from column (1) of Table 4 and reweighting by the fixed-effects regression weights ($\alpha = 0$ in Eq. 6), we obtain a weighted coefficient of 0.069, shown in the bottom row of Table 4. This implies that approximately 1/3 of the change from OLS to FE ($\frac{0.069-0.049}{0.12-0.049}$) is driven by the change in family size weights; with the other 2/3 driven by change in identifying variation. Further, reweighting the FE estimates using the OLS weights ($\alpha = 1$ in Eq. 6) produces a coefficient is 0.083. This implies that the imbalance in family size alone causes the FFE estimate to be 50% higher than the estimates without FE.

### 4.4 Monte Carlo Experiments

We perform a Monte Carlo analysis to examine the properties of our proposed reweighting estimators. We use naturally occuring selection into identification from our PSID application and model treatment effects for three settings, allowing the true ATE to be known. Each setting has a

[26] This variance is computed using the population formula, (dividing by $n_g$), rather than the sample formula (dividing by $n_g - 1$). As with the two-step estimator, these weights can also be multiplied by sample weights.

[27] As Gibbons, Suarez and Urbancic (2018) note, we cannot estimate cluster-robust standard errors in the estimation step of the two-step equation: there are fewer clusters than the sum of the count of fixed effects and covariates. However the standard cluster-robust assumptions imply that the $\delta_g$ are independent of one another. This enables Equation 15.

different model of heterogeneity in treatment, which determines the covariates that the researcher uses to generate the propensity score.

We generate the data for the Monte Carlo as follows: To construct baseline outcomes (i.e. without treatment), we run a linear probability model predicting attainment of "some college or more" with demographic variables, income during childhood, and parental education. From this model we construct a one-dimensional covariate, $X_{ig}$, which is a continuous probability that an individual completes some college.[28] All simulations start with this constructed variable $X_{ig}$ and the variable $HeadStart_{ig}$ from the original data. We then construct latent outcomes inclusive of treatment as $Y_{ig}^{*} = X_{ig} + \beta_{ig} HeadStart$, where $\beta_{ig}$ is the treatment effect of Head Start. We scale $Y_{ig}^{*}$ to ensure that these probabilities lie within the range $[0, 1]$. We then randomly generate the binary outcome variable as $Pr\left(Y_{ig} = 1\right) = Y_{ig}^{*}$.

We consider three models of heterogeneity in treatment effects. First, $\beta_{ig} = 0.08$. We use the variable $X_{ig}$ to generate propensity scores. Second, $\beta_{ig} = 0.192$ for large families (with 4 or more siblings) and $\beta_{ig} = 0$ for small families (3 or fewer children). We use a dummy variable for "large family" to generate propensity scores. Third, we allow the treatment effect heterogeneity to vary smoothly: $\beta_{ig} = 0.08 \cdot \left(1 - \frac{X_{ig} - \bar{X_{ig}}}{s.d.(X_{ig})}\right) \cdot \frac{1}{3}$, with $\bar{X_{ig}}$ and $s.d.\left(X_{ig}\right)$ the mean and standard deviation of $X_{ig}$. This produces a treatment effect that is larger for lower-baseline-probability individuals and ranges from 0.01 to 0.15 for most of the population. For this more complex treatment effect, we generate propensity scores in two ways: using $X_{ig}$ and, more flexibly, using a spline in $X_{ig}$, with knots at the $5^{th}$, $20^{th}$, $50^{th}$, $80^{th}$, and $95^{th}$ percentiles of $X_{ig}$. The latter model presumes that the researcher has some intuition that the treatment effect or selection into identification may vary non-linearly with baseline outcomes.

We run 10,000 replications of our Monte Carlo simulation. In each replication, we keep track of the true ATE for each target population of interest, the FE estimate of the treatment effect, and the reweighted regression estimate of the treatment effect for each target population.[29] The FE estimate is the same for all target populations. We consider four target populations: (i) individuals in Head Start switching families;[30] (ii) all siblings; (iii) all individuals in the sample (including singletons); and (iv) all Head Start participants. We multiply all estimates by 1,000 for easier readability.

Panel A of Table 5 presents results for the model with constant treatment effects. In this setting, the average treatment effect is the same for all target populations, all estimators are unbiased, and the FE model is the minimum variance estimator. The reweighting estimators have mean squared errors 4 to 17% larger than for OLS.

Panel B of Table 5 presents results for the model with zero treatment effect for small families, and large treatment effects for large (4+ children) families. It shows that for every target population,

[28] For simplicity, we restrict the sample to those with $X_{ig} \in [0, 1]$ at baseline.

[29] Both post-regression and in-regression reweighting produce the same results.

[30] This will not necessarily be the same as the FE estimate because of differences in the conditional variance across families.

FE is biased, while the reweighting estimator is always unbiased. This improvement in bias over FE leads to much better mean squared error results for the reweighting estimator.[31]

Panels C and D of Table 5 examines the third model with heterogeneous treatment effect that varies with $X_{ig}$. Here the FE model has relatively little bias for the switcher and Head Start participant targets (-0.11 p.p, and -0.12 p.p. on a base of 9 p.p.), but has 10 times larger bias for the remaining targets. Panel C shows that the regression reweighting estimator which uses $X_{ig}$ in the propensity score estimation has less bias than FE for all target populations, with no detectable bias for the switcher, or Head Start populations. The small bias for the reweighting estimator for the other target populations results from an imperfect balance in the $X_{ig}$ variable, even after reweighting.

Panel D shows that when we re-estimate the model including a spline in $X_{ig}$ in the propensity score estimation, the reweighting estimator has no detectable bias for any of the target groups. This suggests that allowing for greater flexibility in the functional form relationship between covariates and the propensity score can achieve greater reductions in bias.

Overall, the results of this exercise show that that the reweighted estimator has significantly less bias than FE (for the types of treatment effect heterogeneity we consider), and can be successfully targeted toward different target populations. Consistent with the conditioning on observables requirements of this estimator, its performance is best when it is given the appropriate covariates for the particular type of heterogeneity at work.

# 5 Extensions

## 5.1 Unit $i$ Covariates

We now consider FE models that include covariates $C_i$ that vary across $i$ units within a group. Researchers may want to include $C_i$ in their models in order to (i) make Assumption 1 more reasonable; (ii) improve precision of estimates (iii) allow extrapolation to target groups defined at the unit level.

Once these covariates are included, the typical intuition that "groups with variation in treatment" provide identification breaks down. This is because for some groups, who we refer to as "residual switchers," there can be variation in the treatment residualized of $C_i$, even if there is no within-group variation in $D_i$.[32] Thus, treatment effects can also be estimated for residual switchers; however, identifying variation comes from from within-family variation in $C_i$, not $D_i$.

How much do residual switchers matter for estimates? We can quantify this by calculating the share of variation in $D_i$ coming from residual switchers, using a formula similar to the calculation

[31] In results not reported, we have examined adding $X_g$ as a covariate to the propensity score estimation stage in this model. This introduces a small amount of bias in the reweighting estimator (-0.1 p.p., relative to the 2 to 3 p.p. bias in FE) for the "siblings" and "all" target groups.

[32] See Appendix A.3 for a formalization of this.

of the effective number of observations.[33] In our PSID application, residual switchers provide 3% of the variation used for identification of the Head Start FFE coefficient. Therefore, this contributes minimally to the FE estimate.

We can also consider incorporating residual switchers into the reweighting methods. For a general discussion of how our key assumptions and proposition can be extended to accomodate $C_i$, see Appendix A.3. The decision to include residual switchers can vary across contexts, and should depend on the extent to which variation from residual switchers is valid for identifying treatment effects. For example, in our application, residual switchers are primarily families where no children attended Head Start. As a result, we believe that variation from these families is not aligned with our desired thought experiment, which leads us to ignore these families in reweighting. In contrast, in a setting like difference-in-difference, untreated "residual switchers" can provide equal identifying variation as the switchers, which makes it is appropriate to include variation from all groups.

### 5.2 Nonlinear Functional Form

Next, we relax the linear functional form assumption used to demonstrate SI in our Monte Carlo simulations. One reason this may make a difference is that conditional or fixed effect logit and probit models use only "double switchers," families with variation in both the outcome variable and the treatment variable, rather than "switchers". In Appendix E, we show that the biases from SI are similar in the linear probability model and conditional logit, and that the reweighting we propose is equally effective at reducing bias in both cases.

### 5.3 Continuous $D_i$

Finally, while we have focused on the case where $D_i$ is binary, it is worth noting that SI can also be present when $D_i$ is continuous (since $\hat{\delta}_{g,FE}$ is still only estimated for switching families.) It is not clear how frequently this will manifest in practice, however, since groups are more likely to have variation in a continuous covariate. Even so, it may still be worthwhile to verify the number of switchers, since there may be persistent bunching at one value of $D_i$, such as at zero maternal income or at zero instances of an uncommon event.

## 6 Effects of Head Start and Other Applications

### 6.1 Data and Replication of GTC and Deming (2009)

We now turn to examining the impact of Head Start on long run outcomes using the PSID and CNLSY, which were used to analyze this question in GTC and Deming (2009).

[33] In particular, the share of identification from residual switchers is equal to $1 - \frac{\sum_g Var(D_i|C_i, g(i)=g)\cdot(n_g-1)\cdot\mathbf{1}(S_g=1)}{\sum_g Var(D_i|C_i, g(i)=g)\cdot(n_g-1)}$. Alternatively, calculating the effective number of observations using Equation 3 (altering the variance to condition on $C_i$) would produce similar results

### 6.1.1 PSID

The PSID sample includes the sample of individuals surveyed in the PSID by 2011. The PSID began in 1968 as a survey of roughly 5,000 households and has followed the members of these founding households and their children longitudinally. The longitudinal nature of the study allows sibling comparisons during early adulthood as well as later in life.

We begin our analysis with a replication of GTC. The sample includes all black or white individuals born between 1966 and 1977, and excludes Hispanic individuals.We provide a detailed description of our replication of GTC in Appendix D. Despite some minor differences, the two PSID samples are qualitatively similar. The summary statistics are often within a third of a standard deviation of each other. Moreover, the estimated effects of Head Start in this sample are similar to those estimated in GTC. We find large (23 p.p.) and significant effects of Head Start on the probability that whites attain some college, and large point estimates (9.3 p.p.) for high school graduation, though in our case these are not statistically significant. We do not find that Head Start meaningfully reduces the probability of committing a crime.[34]

For the remaining analyses from here, we use a sample that substantially expands and modifies the GTC sample. First, we expand the sample to include individuals born between 1978 and 1987. The individuals in these cohorts were too young when the analysis in GTC was performed to observe their education and early career outcomes. Second, we include older siblings of all individuals, including those born prior to 1966. These early cohorts were typically too old to benefit from the introduction of Head Start, and serve as a plausible control group for the early cohorts.

In addition to modifications of the sample, we also expand the number of outcomes under analysis in order to gain a more extensive understanding of the channels by which Head Start affects children's lives. We follow the established practice of distilling the measures to summary indices to lessen problems with multiple hypothesis testing (see, e.g., Anderson, 2008; Kling, Liebman and Katz, 2007; Hoynes, Schanzenbach and Almond, 2016). We create four indices to capture economic and health outcomes observed for individuals at age 30 and 40. The "economic sufficiency index" includes measures of educational attainment, receipt of AFDC/TANF, food stamps, mean earnings, mean family income relative to the poverty threshold, the fraction of years with positive earnings, the fraction of years that the individual did not report an unemployment spell, and homeownership. The "good health index" summarizes the following component measures: non-smoking, report of good health, and negative of mean BMI.[35]

The process of creating each index follows the procedure described in Kling, Liebman and Katz (2007). In particular, we standardize each component of the index by subtracting the mean outcome for non-treated children, defined as children that did not attend any form of preschool,

[34] In some subsamples, we even find an effect in the opposite direction. We believe these cases are driven by situations where there are rather few observations identifying the coefficients, and that the lack of correspondence may be driven by very minor (and un-diagnosable) differences in specification and/or dataset construction.

[35] See Appendix Table B.4 for descriptive statistics of the inputs to the indices.

and then dividing the result by the standard deviation of the outcome for non-treated children. The summary index takes a mean of these standardized measures.[36] We also extract the first principal component of the standardized variables for "economic sufficiency" and for "good health". Later we use these as alternative outcome variables.

Appendix Table B.2 reports sample descriptive statistics for the expanded sample we construct. For ease of comparison with our earlier replication, we include means for the entire sample, the subsamples of Head Start participants/non-participants, and for the sample of individuals with siblings. We present the means of the analyzed outcomes in Appendix Table B.3.[37]

### 6.1.2 CNLSY

We obtain the CNLSY sample from the Deming (2009) replication files, which ensures that the samples are identical. The CNLSY is a longitudinal survey that follows the children born to the roughly 6,000 women that took part in the NLSY79 survey. The sample we use includes all children who were at least 4 years old by 1990.

## 6.2 Head Start Estimation

The empirical strategy takes advantage of within-family variation in participation in Head Start to identify the long term impact of the program. Following GTC and Deming (2009), we estimate:

$$Y_{ig} = \alpha + \beta_1 HeadStart_{ig} + \beta_2 OtherPreSchool_{ig} + \mathbf{X_{ig}}\gamma + \delta_g + \varepsilon_{ig} \tag{16}$$

where $Y_{ig}$ represents a long-term outcome for individual $i$ with mother $g$. $HeadStart_{ig}$ indicates whether a child reports participation in the program, and $OtherPreSchool_{ig}$ indicates participation in other preschool (and no participation in Head Start). These two variables are in this way defined so as to be mutually exclusive, with "neither Head Start nor other preschool" as the omitted category.[38] $\delta_g$ is a mother fixed effect which enables comparisons across siblings with a shared mother. The vector $X_{ig}$ includes a large number of controls for individual and family characteristics to absorb differences in personal and household characteristics which may be correlated with one's participation in Head Start and long term outcomes. These controls vary due to data availability across sources and specification used in earlier work, but fall into three broad categories: demographics, family background, and family economic circumstances during early childhood.[39]

[36] Consistent with Kling, Liebman and Katz (2007), we generate a summary index for any individual for whom we observe a response for one component of the index. Missing components of the index are imputed as the mean of the outcome conditional on treatment status. For example, if a former Head Start participant is missing an outcome, it is imputed as the mean outcome of other Head Start participants. Likewise for other preschool, or non-preschool participants.

[37] Appendix Table B.4 includes summary statistics for the inputs to the summary indices. Appendix Tables B.5, B.6, and B.7 contain the number of observations for each outcome and control variable in the analysis .

[38] Since Head Start only became available in 1965, we recode Head Start attendance to be "other preschool" for the 1961 and older cohorts.

[39] For the PSID, these include: individual's year of birth, sex, race, and an indicator for being low birth weight,

Missing control variables are imputed at the mean, and we include an indicator variable for these imputed observations. We cluster standard errors on mother id, and use population-representative weights where appropriate.[40] When $Y_{ig}$ is binary, we estimate linear probability models as a main specification and check the sensitivity of our results to alternative models.

The coefficient of interest is $\beta_1$, the impact of Head Start on long term outcomes compared to no preschool. We generate propensity score weights to obtain the ATE for three target populations: (1) Head-Start-eligible individuals, based on family income between ages 2 and 5;[41] (2) all Head Start participants; and (3) all siblings.[42] For parsimony, we use a subset of the variables in Table 2 to generate the propensity score for each race: year of birth, gender, mother's years of education, income at age 3, and income at age 4, and a linear and quadratic in number of siblings. We include results for the post-regression weighting method; results are qualitatively similar when we use in-regression weighting.

### 6.2.1 Evidence on Model Assumptions: Identifying and Conditional Independence

The standard test of the identifying assumption (Assumption 1) is to look for balance in observables across siblings within families. Deming (2009) finds little evidence that Head Start attendance is correlated with observable differences across siblings, which suggests that the magnitude of selection may be small. In Appendix Table B.8, we examine the plausibility of the identifying assumption in the PSID by testing the correlation between participation in Head Start and observable pre-Head Start individual and family characteristics. For the white sample which forms our focus, there are few statistically significant correlations, which suggest that the assumption may be reasonable.[43]

As a test of CFEI, Appendix Table B.9 examines whether treatment effects vary by the share of siblings in the target group. Specifically, we regress estimated family-specific treatment effects on an indicator for whether an individual is a member of the target population, weighting for balance on observables.[44] For most outcomes, this test passes, with no sign of systematic differences across

mother and father's years of education, an indicator for having a single mother at age 4, 4-knot splines in annual family income for each age 0, 1, and 2, a fourth spline based on average family income between ages 3 and 6, indicators for mother's employment status at ages 0, 1, and 2, and household size at age 4. For the CNLSY, these include: health conditions before age 5, PPVT test score at age 3, measures of birth weight, measures of mother's health and health behaviors, mother's working behavior and income prior to age 4, indicator for being first born, participation in Medicaid, relative care, and indicators for early care types.

[40] We follow our predecessors' weighting practices: for the PSID, we generate representative population weights from the 1995 March CPS, and for the CNLSY do not use weights.

[41] An individual is considered Head-Start-eligible if at any point between the ages of 2 and 5 her family income was below 150% of the poverty level, to account for our imperfect ability to observe reportable income.

[42] Propensity score weights are estimated using information on year of birth, maternal education, sex, and maternal income at ages 3 and 4.

[43] For the black sample, participation in Head Start is correlated with a greater likelihood of having higher income at age 1, and lower income at age 2, which may raise concerns that black families may tend to send their children to Head Start after a rupture in the family or after an income shock. However, given the many hypotheses being tested in this table, these significant findings might be spurious; and these results are somewhat sensitive, becoming insignificant when we drop observations with imputed controls.

[44] For target individuals the weights are $1/Pr\left[T_i = 1, S_g = 1\right]$, and for non-target individuals the weights are $1/\Pr[T_i = 0, S_g = 1]$.

target and non-target individuals. However in the CNLSY for the learning disability and poor health outcomes, and the target population of Head Start participants, there is some evidence ($p < 0.10$) of differential treatment effects for HS participants compared to non-participants. Consequently, we advise that the results for these outcomes and target population be viewed with some caution.[45]

Our reweighting procedure also relies on adequate overlap of $Q_x$ across switchers and individuals in the target population in the non-switching sample. In Appendix Figure B.3 we show the density of the estimated probabilities of being a Head Start participant for the switching sample and the non-switching Head Start participant sample. This figure shows that there is a good deal of overlap across the two groups, but also that there are a few Head Start participants whose p-scores lie outside the range of the switchers. These observations represent 5 individuals, 6% of the Head Start non-switcher observations, and 3% of all Head Start participants. We interpret this magnitude of violation of the overlap assumption as mild enough to disregard in our subsequent analysis.[46]

## 6.3 Head Start Results

### 6.3.1 Reweighted Estimates

We begin by presenting results for our illustrative outcome, attainment of some college for whites in the PSID, in Panel A of Table 6. Column (1) of the table presents the estimated impact of Head Start on some college in GTC, column (2) presents the results using our expanded sample, and columns (3) to (5) present reweighted estimates for the three target populations. As reported earlier, we estimate that Head Start increases the likelihood of attaining some college by a statistically significant 12 p.p. (se: 0.053) using the baseline FFE model. This estimate is 57% smaller than the estimate reported in GTC, 0.281 (se: 0.108).[47] The standard errors are also roughly 50% smaller, corresponding to the roughly tripling of sample size (2,986 compared with 1,036).

As we foreshadowed earlier, these estimates are unlikely to represent the ATE for policy relevant populations, such as the Head Start eligible population and Head Start participants. Figure 3 shows a scatter of the FFE weights and the Head-Start-representative weights for each family in the white sample, divided by 2 to 3 child families (Panel A) and 4 or more child families (Panel B). The larger (smaller) markers signify that the estimated effect of Head Start on some college for the family is above (below) median. We also include a 45 degree line for reference. The figure shows that, in general, the Head-Start-representative weights are higher than the FFE weights for small families that experience smaller impacts of Head Start. Conversely, the representative weights are lower

[45] When the target population is Head Start participants, this requirement forces a degree of balance across the target and non-target groups. Another way of viewing this test is: do switching families with a greater share of participants have different coefficients on Head Start than those with a smaller share of participants? We have run analogous models at the family level, which give qualitatively similar results.

[46] We provide the equivalent figure for the "Head-Start-eligible" target population in Appendix Figure B.4. For this target group, the range of switching sample estimated p-scores encompasses that for non-switching target observations.

[47] We show in the appendix that this discrepancy is not due to faulty replication of the GTC estimates in a smaller sample. We estimate a coefficient of 0.232 (se: 0.094) for this sample and outcome in our replication.

relative to the FFE weights for large families that experience larger impacts of Head Start. Hence, we should expect the reweighted estimates to show a reduced impact of Head Start relative to FFE.

The reweighted estimate of the impact of Head Start for the eligible, participant, and sibling populations is between 0.071, 0.031, and 0.075, respectively, and are all statistically insignificant. Setting aside the lack of precision in the estimates, these represent moderately large impacts relative to the 43.7% average rate of college going among Head Start eligible children. But comparing to the FFE coefficient, these effects imply a 38% to 74% smaller impact on college attendance. Putting these estimates in broader perspective, they are 22 to 89% smaller than the unadjusted estimates for *all* participants from other FFE studies (Bauer and Schanzenbach, 2016; Deming, 2009) and 42% smaller than the estimate from the county roll-out of Head Start (Bailey, Sun and Timpe, 2018), although the lower end of the confidence intervals for these estimates include our ATE.

Panel B of Table 6 presents results for the Economic Sufficiency Index in the PSID. Our FFE estimate shows a statistically insignificant 0.023 SD decline in this index associated with Head Start. When we reweight the effects, we find slightly larger negative effects for Head Start eligible children and Head Start participants, and a positive effect (0.03 SD) for siblings. It bears emphasizing, though, that the results are not precisely estimated, such that the 95% confidence intervals allow for a sizeable positive impact of Head Start in spite of the small or negative point estimate. For example, the confidence interval for the economic index for whites allows for a Head-Start-induced improvement of 0.16 SD or a reduction of 0.21 SD for Head Start participants. This limits our ability to make firm conclusions about Head Start's impact on this outcome.

The following four panels of Table 6 show the CNLSY FFE estimates, those reported in Deming (2009) and our replication, and our reweighted estimates. The panels report effects for high school graduation, idleness (not in school or at work), diagnosis of a learning disability, and poor health (based on self-reported health status). The FFE estimates indicate that Head Start leads to a 8.5 p.p. increase in high school graduation ($p < 0.01$), 7.2 p.p. decline in idleness ($p < 0.10$), 5.9 p.p. decline in having a learning disability ($p < 0.01$), and a 6.9 p.p. decline in reporting poor health ($p < 0.01$). The reweighted estimate for participants for high school is 40% smaller, and marginally significant ($p < 0.10$). We also see a substantial 34% decline in the estimated impact on idleness when we consider the impact on participants. The disability and poor health estimates are relatively more stable; the reweighted impacts on participants are just 4% and 25% smaller than the FFE estimate.

In the final column of the table, we test whether the difference between the reweighted estimate for participants and the FFE estimate is statistically significant. We bootstrap the standard errors for this difference by taking draws with replacement from the sample and performing the FFE estimation and reweighting again. We do this 1,000 times and obtain the standard error of our difference as the standard deviation of the 1,000 estimated FFE-reweighted differences. We find that the reweighted estimates for some college (PSID) and high school graduation (CNLSY) are statistically different from the FFE estimate at the 5% and 10% levels, respectively. The remainder

of the outcomes are more imprecisely estimated, and therefore we can not reject that the reweighted estimate is the same as the FFE estimate (in these cases the quantitative differences are non-trivial).

Returning to the PSID, Appendix Tables B.10 and B.11 show the PSID FFE estimates and reweighted results for high school and the good health index for whites, and the corresponding results for blacks. Overall, the results suggest little support for a positive long term effect of Head Start. This is true for the FFE estimates and the reweighted estimates. Nonetheless, the magnitude of the estimates can vary importantly with reweighting, particularly for whites. This makes sense since the identifying sample is a much smaller share of the overall sample for whites relative to blacks. For example, the FFE estimate for the good health index for whites is -0.265 SD, but reweighting for the Head Start participant population changes this estimate to -0.423. In contrast, the coefficients are relatively stable for blacks.[48]

Appendix Table B.12 presents the results when we reweight the FFE estimates using sample shares instead of propensity score weights, which allows us to recover the ATE for the sample of switchers (Gibbons, Suarez and Urbancic, 2018). This estimate is quite similar to the FFE estimate for all outcomes, underscoring that the conditional-variance-weighting plays a minor role in this setting.

### 6.3.2 More Evidence on the Role of Family Size

One key pattern in our findings is that larger families appear to have larger returns to Head Start than smaller families. We believe this to be a new finding in the Head Start literature. We note that this was not a pattern we initially set out to test in this study, so there is some chance of this finding being inadvertently driven by chance and our limited sample sizes. However we think that this may provide an interesting hypothesis for future studies. Also, we first observed this pattern in the PSID data, and so our CNLSY results (see e.g. Columns 3, 4, and 5 of Table 4) are to some degree an out of sample confirmation of this pattern.

We have examined whether the larger coefficients for larger family sizes in Table 4 are driven by family size standing in for other covariates. In Appendix Table B.13 we perform a "horse race" analysis, comparing whether heterogeneous coefficients load on to family size, or other covariates. This table shows that the heterogeneity with family size is robust to also allowing for heterogeity along other covariates. We have also experimented with specifications that test for whether larger family size is merely proxying for "longer sibling cohort span," and do not find evidence that this is the case.

[48] For the black sample, most estimates are also statistically insignificant. However, for the age 30 Economic Sufficiency Index, the reweighted estimates indicate statistically significant negative impacts of Head Start. For example, for a target population of participants the reweighted coefficient on Head Start is -0.208 (s.e. = 0.072).

### 6.3.3 Additional FFE Estimates

Continuing our analysis of the PSID, we also investigate effects of Head Start on a variety of additional short-term outcomes, outcomes at age 40, as well as heterogeneity by race, gender by cohort in Appendix C. We do not find any systematic evidence of effects on any of these outcomes, or important heterogeneity along these dimensions.

## 7 Other Applications

We have shown empirical evidence for selection into identification for three FFE applications relating to the returns to human capital investment and returns to domestic migration. In each of these contexts, there appears to be a mechanical relationship between $Pr(S_g = 1)$ and family size. In the Head Start setting, heterogeneity along these lines creates an upward-bias in the FE estimate. Since returns to migration may also be heterogeneous by family characteristics, it may be useful to also reweight the estimates from Collins and Wanamaker (2014) to obtain the ATE for a representative set of migrants.

We now discuss three additional FE designs present in the education, labor, and environmental literatures that illustrate settings where the tools that we have developed may apply. First, a number of studies examine the effect of peers in the classroom within a school-grade (or school) using school-grade FE (or school FE). For example, Carrell and Hoekstra (2010) and Carrell, Hoekstra and Kuka (2018) examine the effect of having a peer exposed to domestic violence (DV) using this strategy, finding large negative impacts on contemporaneous achievement that persist to reduce long-term earnings. While the DV measure in these studies is continuous, it is reasonable to think that this may be a "lumpy" variable in the sense that some schools (or school-grades), which have a low probability of DV, will never have a student exposed to DV during the 8 year window of observation, and some school-grades, which have a high probability of DV, may always have a student exposed to DV. Given the likely correlation between $Pr(DV_g)$ and $Pr(S_g = 1)$, non-switcher schools probably also have a different set of school resources (e.g. share of highly experienced teachers) and student composition (e.g. mean family income) than switchers, which could either exacerbate or mitigate the effects of DV. As a result, the effects estimated from switching schools may not generalize to low-probability-DV non-switchers or high-probability-DV non-switchers.

Second, a set of influential papers by Dube, Lester and Reich (2010, 2015) identify the impact of minimum wage laws within border-county pairs (using border-pair-by-year FE). This strategy produces bounds on minimum wage elasticities which are less negative than those estimated with other strategies. The authors report that 91% of the county pairs in the data have variation in the minimum wage at some point during the analysis, however states with more border counties and who have more frequent changes to the minimum wage relative to neighboring states will contribute more variation to the design. Hence, in practice, identification may be concentrated

among a subset of the 91%. At the same time, the characteristics of switching border counties are likely to be different from interior counties, in terms of the education distribution, population density, or industry composition, which could influence the response to minimum wage increases (Cengiz et al., 2019). Thus, reweighting the estimates of switching border pairs to account for these characteristics could yield a different estimate for the impact of the minimum wage.

Third, it has become common to estimate the effect of environmental shocks on health and human capital using variation in temperature or rainfall within a local area (e.g. district FE). For example, Shah and Steinberg (2017) employ this strategy and find that a positive rain shock (top 20% rainfall) reduces the likelihood that students attend school, and vice versa for droughts. Since shocks are by definition infrequent events, it is likely that some districts that have more moderate climates will have no shocks over the 4 years of analysis. These non-switching districts may be located in a different geography, have distinct industrial composition, or population characteristics, which could in turn affect the elasticity of school attendance. Hence, extrapolating from switcher to non-switcher districts may require reweighting strategies such as those we propose.

These applications highlight the fact that selection into identification is likely to be relevant across the numerous domains where FE are applied. We leave it to future researchers to quantify the role of this selection and apply reweighting techniques to test the sensitivity of the conclusions.

# 8 Conclusion

Fixed effects can provide a useful approach for treatment effect estimation. The *internal* validity of this strategy, which has been the subject of much debate, relies on the assumption that treatment is randomly assigned to units in a group. In this article, we show that an additional assumption is needed for the *external* validity of results: that groups with variation (switchers) have comparable treatment effects to groups without variation (non-switchers). In other words, fixed effects estimates are generalizable only if there is no *selection into identification*.

We show that this assumption is not trivial in the context of family fixed effects. We document across multiple settings that switching families are systematically larger and show that this can induce bias in estimation. We develop a novel approach to recover ATE's for representative populations, which upweights observations that are under-represented in the identifying sample relative to the population of interest. We demonstrate that this reweighting approach performs well using Monte Carlo simulations.

We apply these lessons to an analysis of the long term effects of Head Start in the PSID and CNLSY using family fixed effects. Relative to prior evaluations of Head Start using FFE in the PSID, we use a sample three times as large in size, include longer run (up to age 40) outcomes, and expand the set of outcomes under consideration. Echoing prior findings, we find using FFE that Head Start significantly increases the likelihood of completing some college and graduating from high school, and decreases the likelihood of being idle, having a disability, or reporting poor health.

Using our reweighting methods, we estimate that Head Start leads to a 3.1 p.p. increase in the likelihood of attending some college for Head Start participants, and a 7.1 p.p. increase for Head Start eligible. The ATE estimate for participants is 70% smaller than the FFE estimate, a difference which is statistically significant at the 5% level. We examine several other outcomes and find few statistically significant results. In sum, the FFE results in the PSID indicate that Head Start has little effect on many long term outcomes on average, with the exception of completing some college, and perhaps even detrimental effects for men. In the CNLSY, for high school graduation we find that the reweighted estimate for participants (5.1 p.p.) is 40% smaller than the FFE estimate, a difference which is statistically different at the 10% level. We find less change associated with reweighting for other outcomes.

Overall, we interpret our findings as pointing primarily toward "increased uncertainty" and to a limited degree toward "zero effects" of the Head Start program. This suggests that there is some discordance between the long-term results from the FFE design, and new estimates using other designs, which generally produce larger and more robust effects of this intervention. Reconciling these findings is beyond the scope of this paper, but would be a productive avenue for future work.

Based on our findings, we propose new standards for practice when using FE or similar research designs to diagnose, and potentially correct for, the role of changes in sample composition in explaining the gap between OLS and FE estimates.

1. First, analyses should report the switching sample size in addition to the total sample size, including for relevant subsamples of the data (e.g. whites and blacks). It may also be useful to calculate the effective number of observations and share of identifying variation from true switchers to increase transparency into the variation among switchers.

2. Second, we suggest that researchers show a balance of observables across switching status to complement evidence of within-sample balance across treatment status. These covariates should include the number of units in a group (if there is imbalance) and correlates of treatment. For example, in the case of movers, one might consider testing for balance of urbanicity, age, and occupations. If there are differences in these covariates, researchers should examine heterogeneity along these dimensions. These tests are likely to have limited power to detect issues if there are interactions between covariates, but are a useful bellweather for important external validity concerns.

3. As a subsequent step, we recommend using propensity-score reweighting of the FE estimates to obtain estimates for a representative population or a policy-relevant population, such as program participants. Since these methods can perform unevenly under some models of heterogeneity, we suggest testing for sensitivity of results and reporting a range of estimates where applicable.

## References


**Abrevaya, Jason.** 2006. "Estimating the effect of smoking on birth outcomes using a matched panel data approach." *Journal of Applied Econometrics*, 21(4): 489–519.

**Aizer, Anna, and Flavio Cunha.** 2012. "The Production of Human Capital: Endowments, Investments and Fertility." National Bureau of Economic Research Working Paper 18429. DOI: 10.3386/w18429.

**Almond, Douglas, Kenneth Y. Chay, and David S. Lee.** 2005. "The Costs of Low Birth Weight." *The Quarterly Journal of Economics*, 120(3): 1031–1083.

**Anderson, Michael L.** 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*, 103.

**Andersson, Fredrik, John C. Haltiwanger, Mark J. Kutzbach, Giordano E. Palloni, Henry O. Pollakowski, and Daniel H. Weinberg.** 2016. "Childhood Housing and Adult Earnings: A Between-Siblings Analysis of Housing Vouchers and Public Housing." National Bureau of Economic Research Working Paper 22721.

**Andrews, Isaiah, and Emily Oster.** 2019. "A Simple Approximation for Evaluating External Validity Bias." *Economics Letters*, 178: 58–62. Working Paper.

**Angrist, Joshua, and Jorn-Steffen Pischke.** 2009. *Mostly Harmless Econometrics.* Princeton University Press.

**Angrist, Joshua D.** 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica*, 66(2): 249–288.

**Angrist, Joshua D., and Ivan Fernandez-Val.** 2013. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." *Advances in Economics and Econometrics: Tenth World Congress*, , ed. Daron Acemoglu, Manuel Arellano and Eddie Dekel Vol. 3 of *Econometric Society Monographs*, 401 – 434. Cambridge University Press.

**Aronow, Peter M., and Allison Carnegie.** 2013. "Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable." *Political Analysis*, 21(4): 492–506.

**Bailey, Martha J., Shuqiao Sun, and Brenden Timpe.** 2018. "Prep School for Poor Kids: The Long-Run Impacts of Head Start on Human Capital and Economic Self-Sufficiency." Working Paper.

**Barr, Andrew, and Chloe R. Gibbs.** 2018. "Breaking the Cycle? Intergenerational Effects of an Anti-Poverty Program in Early Childhood." *mimeo.*

**Bauer, Lauren, and Diane Whitmore Schanzenbach.** 2016. "The Long-Term Impact of the Head Start Program." *The Hamilton Project.*

**Bayer, Patrick, Randi Hjalmarsson, and David Pozen.** 2009. "Building Criminal Capital behind Bars: Peer Effects in Juvenile Corrections*." *The Quarterly Journal of Economics*, 124(1): 105–147.

**Beck, Nathaniel.** 2015. "Estimating Grouped Data Models with a Binary Dependent Variable and Fixed Effects: What are the Issues? Comments Prepared for Delivery at the Annual Meeting of the Society for Political Methodology." *mimeo.*

**Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes.** 2007. "From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes*." *The Quarterly Journal of Economics*, 122(1): 409–439.

**Borusyak, Kirill, and Xavier Jaravel.** 2017. "Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume." Working Paper.

**Bound, John, and Gary Solon.** 1999. "Double Trouble: On the Value of Twins-based Estimation of the Return to Schooling." *Economics of Education Review*, 18(2): 169–182.

**Callaway, Brantly, and Pedro H. C. Sant'Anna.** 2018. "Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment." Working Paper.

**Cameron, A. Colin, and Pravin K. Trivedi.** 2005. *Microeconometrics: Methods and Applications.* Cambridge University Press.

**Carneiro, Pedro, and Rita Ginja.** 2014. "Long-Term Impacts of Compensatory Preschool on Health and Behavior: Evidence from Head Start." *American Economic Journal: Economic Policy*, 6(4): 135–173.

**Carrell, Scott E., and Mark L. Hoekstra.** 2010. "Externalities in the Classroom: How Children Exposed to Domestic Violence Affect Everyone's Kids." *American Economic Journal: Applied Economics*, 2(1): 211–228.

**Carrell, Scott E., Mark Hoekstra, and Elira Kuka.** 2018. "The Long-Run Effects of Disruptive Peers." *American Economic Review*, 108(11): 3377–3415.

**Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer.** 2019. "The Effect of Minimum Wages on Low-Wage Jobs*." *The Quarterly Journal of Economics.*

**Chamberlain, Gary.** 1980. "Analysis of Covariance with Qualitative Data." *The Review of Economic Studies*, 47(1): 225–238.

**Chetty, Raj, and Nathaniel Hendren.** 2018*a*. "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects*." *The Quarterly Journal of Economics*, 133(3): 1107–1162.

**Chetty, Raj, and Nathaniel Hendren.** 2018*b*. "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates." *The Quarterly Journal of Economics*, 133(3): 1163–1228.

**Chorniy, Anna V, Janet Currie, and Lyudmyla Sonchak.** 2018. "Does Prenatal WIC Participation Improve Child Outcomes?" National Bureau of Economic Research Working Paper 24691.

**Collins, William J., and Marianne H. Wanamaker.** 2014. "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data." *American Economic Journal: Applied Economics*, 6(1): 220–252.

**Currie, Janet, and Duncan Thomas.** 1995. "Does Head Start Make a Difference?" *American Economic Review*, 85(3): 341–364.

**Currie, Janet, and Ishita Rajani.** 2015. "Within-Mother Estimates of the Effects of WIC on Birth Outcomes in New York City." *Economic Inquiry*, 53(4): 1691–1701.

**Currie, Janet, and Maya Rossin-Slater.** 2013. "Weathering the storm: Hurricanes and birth outcomes." *Journal of Health Economics*, 32(3): 487 – 503.

**Deming, David.** 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics*, 1(3): 111–134.

**Dube, Arindrajit, T. William Lester, and Michael Reich.** 2010. "Minimum Wage Effects Across State Borders: Estimates Using Contiguous Counties." *The Review of Economics and Statistics*, 92(4): 945–964.

**Dube, Arindrajit, T. William Lester, and Michael Reich.** 2015. "Minimum Wage Shocks, Employment Flows, and Labor Market Frictions." *Journal of Labor Economics*, 34(3): 663–704.

**Fernandez-Val, Ivan.** 2009. "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models." *Journal of Econometrics*, 150(1): 71 – 85.

**Figlio, David, Jonathan Guryan, Krzysztof Karbownik, and Jeffrey Roth.** 2014. "The Effects of Poor Neonatal Health on Children's Cognitive Development." *American Economic Review*, 104(12): 3921–3955.

**Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams.** 2016. "Sources of Geographic Variation in Health Care: Evidence From Patient Migration*." *The Quarterly Journal of Economics*, 131(4): 1681–1726.

**Garces, Eliana, Duncan Thomas, and Janet Currie.** 2002. "Longer-Term Effects of Head Start." *American Economic Review*, 92(4): 999–1012.

**Gibbons, Charles E., Serrato Juan Carlos Suarez, and Michael B. Urbancic.** 2018. "Broken or Fixed Effects?" *Journal of Econometric Methods*, 0(0).

**Gibbs, Chloe, Jens Ludwig, and Douglas L Miller.** 2013. "Does Head Start Do Any Lasting Good?" *Legacies of the War on Poverty*, , ed. Martha J. Bailey and Sheldon Danziger. Russell Sage Foundation.

**Goodman-Bacon, Andrew.** 2018. "Difference-in-Differences with Variation in Treatment Timing." National Bureau of Economic Research Working Paper 25018.

**Hoxby, Caroline M.** 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics*, 115(4): 1239–1285.

**Hoynes, Hilary, Diane Whitmore Schanzenbach, and Douglas Almond.** 2016. "Long-Run Impacts of Childhood Access to the Safety Net." *American Economic Review*, 106(4): 903–934.

**Johnson, Rucker C., and C. Kirabo Jackson.** 2017. "Reducing Inequality Through Dynamic Complementarity: Evidence from Head Start and Public School Spending." National Bureau of Economic Research Working Paper 23489. DOI: 10.3386/w23489.

**Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz.** 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica*, 75(1): 83–119.

**Lemieux, Thomas.** 1998. "Estimating the Effects of Unions on Wage Inequality in a Panel Data Model with Comparative Advantage and Nonrandom Selection." *Journal of Labor Economics*, 16(2): 261–291.

**Lochner, Lance, and Enrico Moretti.** 2015. "Estimating and Testing Models with Many Treatment Levels and Limited Instruments." *The Review of Economics and Statistics*, 97(2): 387–397.

**Loken, Katrine V., Magne Mogstad, and Matthew Wiswall.** 2012. "What Linear Estimators Miss: The Effects of Family Income on Child Outcomes." *American Economic Journal: Applied Economics*, 4(2): 1–35.

**Ludwig, Jens, and Douglas L. Miller.** 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." *The Quarterly Journal of Economics*, 122(1): 159–208.

**Mundlak, Yair.** 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica*, 46(1): 69–85.

**Pages, Remy J.-C., Dylan J. Lukes, Drew H. Bailey, and Greg J. Duncan.** 2019. "Elusive Longer-Run Impacts of Head Start: Replications Within and Across Cohorts." *arXiv:1903.01954 [econ, q-fin]*. arXiv: 1903.01954.

**Rossin-Slater, Maya.** 2013. "WIC in your neighborhood: New evidence on the impacts of geographic access to clinics." *Journal of Public Economics*, 102: 51–69.

**Roy, A. D.** 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers*, 3(2): 135–146.

**Shah, Manisha, and Bryce Millett Steinberg.** 2017. "Drought of Opportunities: Contemporaneous and Long-Term Impacts of Rainfall Shocks on Human Capital." *Journal of Political Economy*, 125(2): 527–561.

**Sloczynski, Tymon.** 2018. "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands." Working Paper.

**Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf.** 2011. "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2): 369–386.

**Suri, Tavneet.** 2011. "Selection and Comparative Advantage in Technology Adoption." *Econometrica*, 79(1): 159–209.

**Thompson, Owen.** 2017. "Head Start's Long-Run Impact: Evidence from the Program's Introduction." *Journal of Human Resources*.

**Verdier, Valentin, and Andrew Castro.** 2019. "Average Treatment Effects for Stayers with Correlated Random Coefficient Models of Panel Data." *mimeo*.

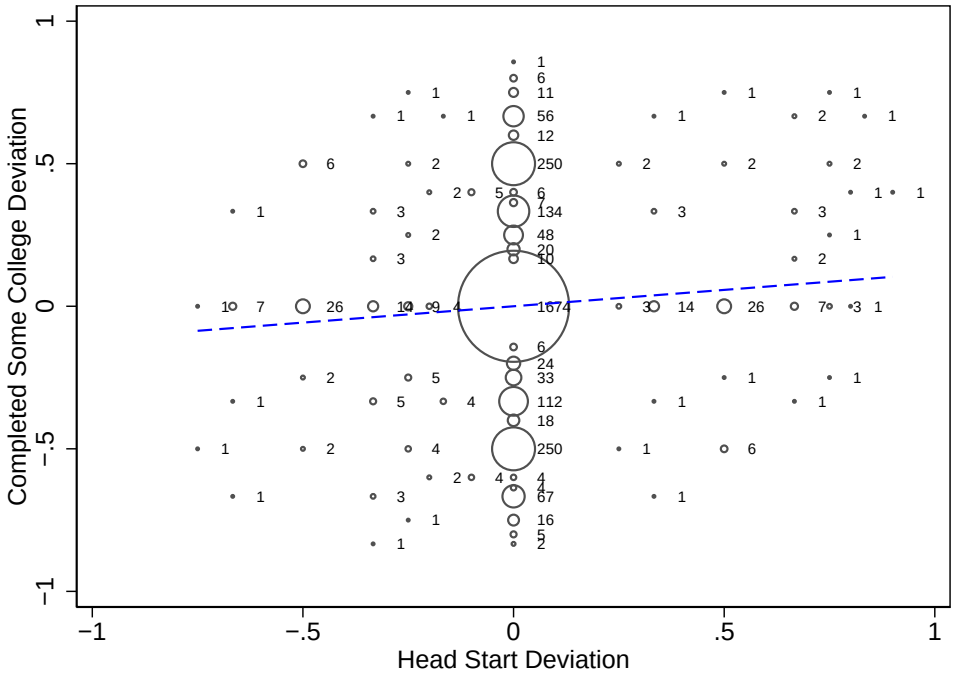
**Wiswall, Matthew.** 2013. "The dynamics of teacher quality." *Journal of Public Economics*, 100: 61–78.

**Wooldridge, Jeffrey M.** 2010. *Econometric Analysis of Cross Section and Panel Data.* MIT Press.

**Xie, Zong-Xian, Shin-Yi Chou, and Jin-Tan Liu.** 2016. "The Short-Run and Long-Run Effects of Birth Weight: Evidence from Large Samples of Siblings and Twins in Taiwan." *Health Economics*, 26(7): 910–921.

# 9 Figures

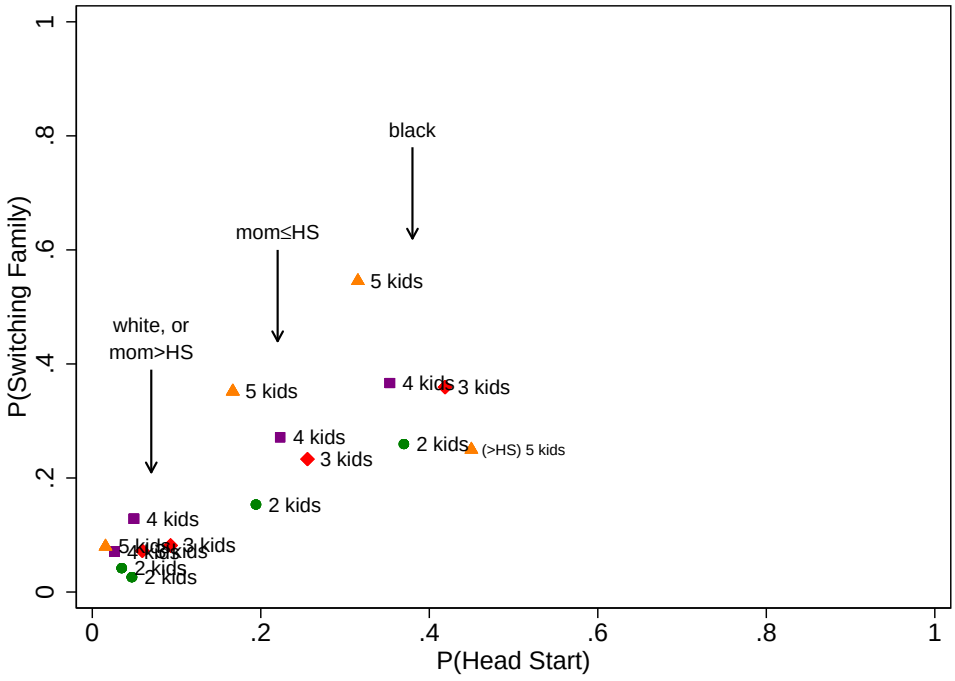
Figure 1: Within-Family Variation in Head Start and Attendance of Some College (PSID)



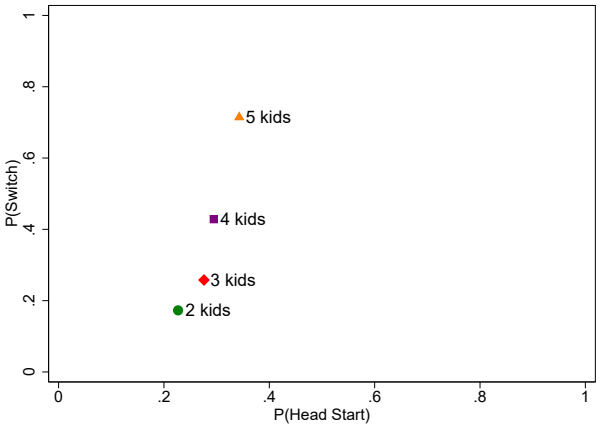

Notes: This figure depicts the identifying variation used in a FFE regression of some college on an indicator for participation in Head Start. Each marker represents the number of individuals that exhibit a particular deviation from the mean Head Start attendance of their family and from the mean attendance of some college of their family. Deviations are defined as the difference between individual attendance of Head Start/some college (1 or 0) and mean of Head Start/some college of one's family. The marker size represents the unweighted number of individuals. We also include a best-fit line, weighted by the number of individuals in each marker. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Figure 2: Likelihood of Being a Switcher Family Increases with Family Size and P(treatment)

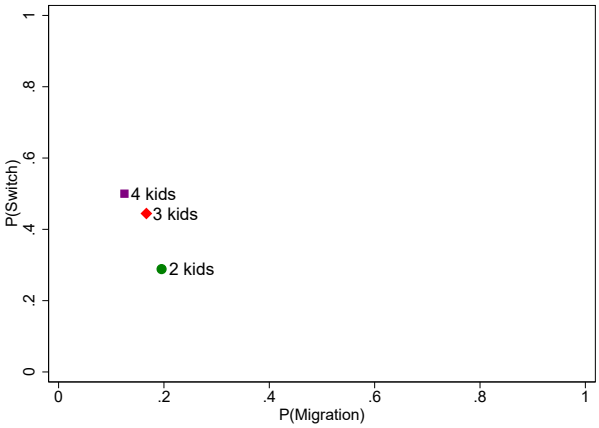
(a) Head Start in PSID




(b) Head Start in CNLSY



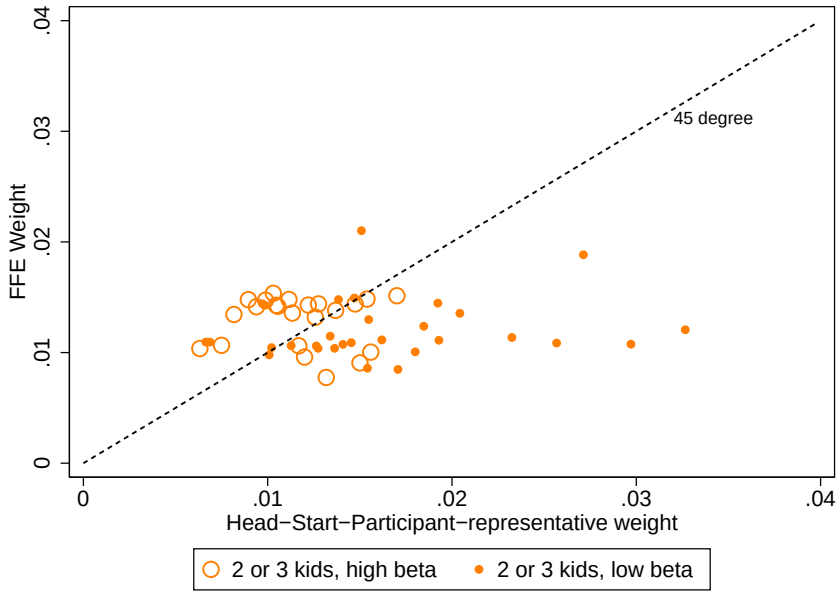

(c) Migration in Census



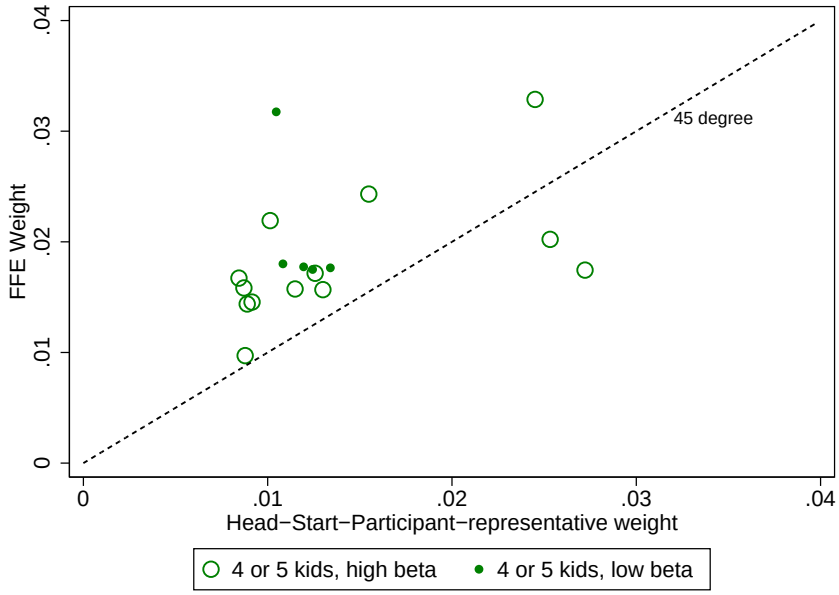

Notes: Panel (a) of this figure plots the *observed* probability of being in a switching family and of attending Head Start by family size for the following groups in the PSID: Whites, Blacks, children of mothers with at most a high school degree, and children of mothers with at least some college. Figure (b) plots analogous markers using data on Head Start participation from the CNLSY. Figure (c) plots the analogous figure substituting migration for Head Start attendance, from a linking of the 1910 to 1930 censuses used in the analysis and made available from Collins and Wanamaker (2014).

Figure 3: FFE Weights and Head-Start-Participant-Representative Weights by Family Size and Some College $\beta$ (PSID White Sample)

(a) Families with 2-3 Children




(b) Families with 4+ Children




Notes: Each marker in this figure indicates the FFE weights and Head-Start-participant-representative (post-regression) weight for one white switching family. The color of the marker indicates whether the family has 2-3 children or 4 or more children. The size of the marker indicates the estimated family-specific beta from a regression of attainment of some college on interactions between Head Start and family id fixed effects. A larger marker indicates an above median beta, while a smaller marker indicates a below-median beta. The 45 degree line is included for reference. Observations above (below) the line are overweighted (underweighted) in the FFE sample relative to a representative Head Start sample. Source: Panel Study of Income Dynamics, 1968-2011 waves.

# 10 Tables

Table 1: Family FE Articles in Top Applied Journals, 2002 to 2017

| | Binary Indep. | Binary Dep. | Both Binary | Total |
|---|---|---|---|---|
| AEJ: Applied | 6 | 4 | 3 | 8 |
| AEJ: Economic Policy | 1 | 1 | 1 | 1 |
| AER | 3 | 1 | 1 | 5 |
| AER Papers and Proceedings | 2 | 2 | 1 | 3 |
| Journal of Health Economics | 5 | 3 | 2 | 7 |
| Journal of Human Resources | 7 | 2 | 2 | 12 |
| Journal of Labor Economics | 2 | 1 | 1 | 5 |
| Journal of Political Economy | 2 | 1 | 1 | 2 |
| Journal of Public Economics | 4 | 4 | 4 | 5 |
| QJE | 1 | 4 | 1 | 4 |
| Review of Economics and Statistics | 2 | 0 | 0 | 3 |
| Total | 35 | 23 | 17 | 55 |
| *Common Dependent Variables* | | | | |
| Schooling/Attainment | 23 | | | |
| Test Score | 17 | | | |
| Employment/Earnings | 15 | | | |
| Birth Weight | 6 | | | |
| Health | 6 | | | |
| Behavioral Issues/Crime | 5 | | | |
| *Common Independent Variables* | | | | |
| Schooling | 8 | | | |
| Birth Weight | 5 | | | |
| Health | 5 | | | |
| Parental Traits | 4 | | | |
| Employment | 3 | | | |
| Birth order | 3 | | | |
| Means-Tested Public Program | 2 | | | |
| Death of Family Member | 2 | | | |
| Bombing/Radiation | 2 | | | |
| *Observations by Sample* | | | | |
| | Siblings N | Total N | | |
| p10 | 469 | 1,212 | | |
| p25 | 1,167 | 2,142 | | |
| p50 | 6,315 | 17,501 | | |
| p75 | 160,122 | 551,630 | | |
| p90 | 750,697 | 1,582,142 | | |
| Year Publication Min/Max | 2002 | 2017 | | |
| Articles with Balance Table if Binary Ind. | 1 | | | |

Notes: This table presents a summary of FFE articles published between January 2000 and May 2017 in 11 top applied journals, which are listed in the first panel of the table. For reference, between 2002 and 2017 the number of articles published in AEJ: Applied was 310; AEJ: Policy was 313; AER was 1722; AER P&P was 1676; JoLE was 434; Journal of Political Economy was 548; QJE was 639; JHR was 543; JPubE was 1688; REStat was 1033; JHE was 1017. Articles were initially idenitified using the search terms "family," "within family," "sibling," "twin," "mother," "father," "brother," "sister," fixed effect," "fixed-effect," and "birthweight" using queries on journal websites. Siblings N is the number of observations reported for the sample of siblings, while Total N represents the number of total observations reported. See text for details.

Table 2: Switchers and Non-Switchers Vary Along Dimensions Other Than Family Size

| | (1) Switch | (2) Non-Switch | (3) T-Stat. (1)=(2) | (4) Beta Switch | (5) T-Stat (4) |
|---|---|---|---|---|---|
| *A. Individual Covariates* | | | | | |
| Fraction female | 0.562 | 0.495 | 4.067 | 0.024 | 0.719 |
| Fraction African-American | 0.516 | 0.111 | 25.877 | 0.249 | 5.640 |
| Mother's yrs education | 9.283 | 11.230 | -21.590 | -0.140 | -0.751 |
| Father's yrs education | 9.190 | 11.371 | -19.594 | -0.389 | -1.784 |
| Had a single mother at age 4 | 0.252 | 0.099 | 10.049 | 0.055 | 2.543 |
| Family income (age 3-6) (CPI adjusted) | 31809 | 52574 | -24.735 | -4759 | -5.719 |
| Mother employed, age 0 | 0.508 | 0.570 | -3.099 | 0.055 | 2.339 |
| Mother employed, age 1 | 0.517 | 0.543 | -1.342 | 0.058 | 2.359 |
| Mother employed, age 2 | 0.536 | 0.554 | -0.951 | 0.118 | 3.565 |
| Household size at age 4 | 5.487 | 4.451 | 12.343 | 0.755 | 4.936 |
| Fraction low birth weight | 0.077 | 0.058 | 1.971 | 0.010 | 0.702 |
| Observations | 1103 | 5500 | 6603 | 7372 | 7372 |
| *B. Inverse Selection into Identification Wts.* | | | | | |
| Pr(switch)/Pr(Head Start), Whites | 2.976 | 2.318 | | | |
| | (1.99) | (1.98) | | | |
| Pr(switch)/Pr(Head Start), Blacks | 1.987 | 1.148 | | | |
| | (1.21) | (1.10) | | | |

Notes: Panel A of this table presents comparisons of the characteristics of individuals in switching families and non-switching families. Columns 1, 2, and 3, respectively, show the mean characteristics of individuals in families that are switchers; individuals in families that are not switchers; and individuals that attended Head Start (HS) in non-switcher families. Column 3 presents the t-statistic for the test that columns 1 and 2 are equal. Column 4 shows the estimates from a regression of each row heading on an indicator for being in a switcher family, with the corresponding t-statistic shown in Column 5, with standard errors clustered on id1968. All controls from the main specification are included excluding the variable shown in the row heading. All estimates are weighted to be representative of 1995 population; see text for details. Panel B shows the mean and standard deviation (in parenthesis) of the inverse of the post-regression propensity score weights when the target is Head Start participants. Pr(switch) and Pr(Head Start) are obtained from a multinomial logit model as described in the text. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table 3: Change in Weighting of Regression Estimates Across Sibling and Switcher Samples (PSID)

| | Number of Children in Family: | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 + |
| *A. Share of Sample* | | | | | |
| All Sample | 0.123 | 0.273 | 0.238 | 0.147 | 0.134 |
| Siblings Sample | 0.000 | 0.345 | 0.300 | 0.186 | 0.169 |
| Switchers Sample | 0.000 | 0.210 | 0.271 | 0.197 | 0.322 |
| *B. Variance in Head Start* | | | | | |
| All Sample | 0.089 | 0.104 | 0.121 | 0.127 | 0.132 |
| Siblings Sample | 0.000 | 0.024 | 0.050 | 0.059 | 0.068 |
| Switchers Sample | 0.000 | 0.045 | 0.098 | 0.131 | 0.174 |
| *C. Regression weights* | | | | | |
| All Sample | 0.171 | 0.257 | 0.284 | 0.117 | 0.101 |
| Siblings Sample | 0.000 | 0.338 | 0.374 | 0.154 | 0.134 |
| Switchers Sample | 0.000 | 0.256 | 0.307 | 0.190 | 0.248 |

Notes: This table shows the change in the composition of the PSID sample moving from all individuals ("All Sample") to individuals that have at least one other sibling in the sample ("Siblings Sample") to individuals in families that have variation in Head Start attendance ("Switchers sample.") Panel A shows the share of individuals in each sample that come from a family with 1 child (zero siblings), 2 children, etc. Panel B shows the variance in Head Start for each family size and sample. For switchers, this is calculated net of family fixed effects. Panel C shows the "regression weight" given to each family size in a given sample, denoted as $\omega_z$ and defined formally in Section 3. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table 4: Returns to Head Start by Family Size,

and Implications for Regression Estimates

| | PSID | | CNLSY | | |
|---|---|---|---|---|---|
| | Some College | | HS Grad | Idle | Lrn. Disab. |
| | CX | FE | FE | FE | FE |
| | (1) | (2) | (3) | (4) | (5) |
| *A. Effects by Family Size* | | | | | |
| Head Start x 1 child family | 0.169* | | | | |
| | (0.091) | | | | |
| Head Start x 2 child family | 0.038 | -0.126 | 0.033 | -0.067 | -0.028 |
| | (0.079) | (0.099) | (0.042) | (0.052) | (0.025) |
| Head Start x 3 child family | -0.030 | 0.152** | 0.061 | -0.038 | -0.070 |
| | (0.087) | (0.075) | (0.060) | (0.068) | (0.043) |
| Head Start x 4 child family | -0.053 | 0.251*** | 0.156* | -0.002 | -0.064 |
| | (0.100) | (0.091) | (0.086) | (0.111) | (0.049) |
| Head Start x 5+ child family | 0.572*** | 0.348*** | 0.277*** | -0.306** | -0.157* |
| | (0.119) | (0.126) | (0.097) | (0.139) | (0.081) |
| Head Start x Unknown child family | -0.099 | | | | |
| | (0.108) | | | | |
| Observations | 4258 | 2986 | 1251 | 1251 | 1247 |
| Head Start Switchers | | 213 | 581 | 581 | 581 |
| Effective Obs. (Indivs. 2-Person Fams) | | 235.9 | 647.9 | 647.9 | 647.9 |
| Effective Obs. (CX Indivs.) | | 731.8 | 438.7 | 438.7 | 438.7 |
| *B. Simulated Estimates across Samples using Family-Size Regression Weights* | | | | | |
| All | 0.046 | | | | |
| Siblings | 0.037 | 0.083 | 0.074 | -0.068 | -0.053 |
| Switchers | 0.069 | 0.123 | 0.088 | -0.073 | -0.060 |

Notes: Panel A of this table shows the coefficients from a regression of some college on a series of indicators for whether an individual attended Head Start interacted with an indicator for the number of children in one's family. The sample is composed of white individuals. Columns 1 include controls, but not mother f.e., and standard errors are clustered at 1968 family id. Column 2 includes mother fixed effects, and standard errors clustered by mother id. The number of Head Start switchers is equal to the number of individuals in families that have variation in Head Start. "Effective Obs. (CX Indivs.)" is the equivalent number of cross-sectional units that provide the same amount of variation as switchers. "Effective Obs. (Indivs. 2-Person Fams)" is the equivalent number of individuals in 2-person switching families that provide the same amount of variation as switchers. Both of these are calculated using Equation 3, where the denominator is the variance of Head Start, residualized by the family mean of the covariates in the analysis, or 0.125, respectively.Panel B shows the weighted average of the coefficients when using regression weights, $\omega_z$ (defined in Section 3), determined by the overall distribution of families ("All"), the distribution of 2+ child families ("Siblings"), and the distribution of 2+ child families that have variation in Head Start attendance ("Switchers"). * p < .10, ** p < .05, *** p < .01. Source: Panel Study of Income Dynamics, 1968-2011 waves and Children of the National Longitudinal Study of Youth.

Table 5: Monte Carlo Experiments: Bias of Reweighting and FFE Relative to True ATE, and Efficiency of Reweighting Relative to FFE

| | | Bias: | | MSE of Reweight |
|---|---|---|---|---|
| | True ATE | FE | Reweight | MSE of FE |
| *A. Constant TE; p-score: $X_g$* | | | | |
| Switchers | 80 | -0.1 | -0.2 | 1.04 |
| Siblings | 80 | -0.1 | -0.0 | 1.16 |
| All | 80 | -0.1 | -0.2 | 1.17 |
| HS Participants | 80 | -0.2 | 0.4 | 1.03 |
| *B. Large family TE; p-score: large family* | | | | |
| Switchers | 83.0 | -10.6* | 0.1 | 0.93 |
| Siblings | 49.8 | 22.6* | -0.0 | 0.68 |
| All | 40.4 | 32.0* | -0.1 | 0.52 |
| HS Participants | 41.2 | 31.3* | -0.1 | 0.53 |
| *C. TE linear in $X_{ig}$; p-score: $X_{ig}$* | | | | |
| Switchers | 93.8 | -1.1* | -0.2 | 1.04 |
| Siblings | 80.2 | 12.5* | 3.2* | 0.98 |
| All | 80.0 | 12.7* | 3.2* | 0.98 |
| HS Participants | 91.5 | 1.2* | 0.3 | 1.03 |
| *D. TE linear in $X_{ig}$; p-score: $X_{ig}$ spline* | | | | |
| Switchers | 93.8 | -1.1* | 0.5 | 1.04 |
| Siblings | 80.2 | 12.5* | 0.6 | 1.09 |
| All | 80.0 | 12.7* | 0.6 | 1.10 |
| HS Participants | 91.5 | 1.2* | 0.5 | 1.09 |

Notes: This table shows the results from 10,000 Monte Carlo simulations. Each panel of the table shows results from a different DGP and/or different covariates used in the p-score, and each row within panel is for a different target population. The true DGP is linear, and is discussed in Section 4.4. The first panel shows results where Head Start has a constant treatment effect (TE) for all individuals; the second shows results where Head Start (HS) has no effect on individuals from small families (3 or fewer children) and a large effect for families with many children (4 or more children); and the third and fourth panels show results where treatment effects that are linear in $X_{ig}$. Column 1, "True Beta," presents the true average increase in the probability of completing some college for participants in Head Start in the sample, which is a function of the DGP and sample composition. Columns 2 and 3 present the bias of various estimation strategies, defined as the difference between the estimated effects of Head Start and the true beta. The estimated effects come from a LPM, propensity-score weighted LPM, respectively. Column 4 presents the ratio of the mean squared error (MSE) of the reweighting estimators relative to LPM. Reweighted estimates are obtained using in-regression weighting, with weights adjusting for the representativeness of switchers and the conditional variance of Head Start within families. All betas are multiplied by 1,000. * $p < .01$.

Table 6: Head Start Impact for Representative Eligible Children, Participants, and Siblings

Using Post-Regression Reweighting Method

| | FFE | | Reweighted ATE, Target = | | | Diff. b/w |
|---|---|---|---|---|---|---|
| | GTC/Deming | Expand Sample/ Replicate | HS Eligible | Participants | Siblings | FFE and Participant ATE |
| *A. Some College (PSID)* | | | | | | |
| Head Start | 0.281** | 0.120** | 0.071 | 0.031 | 0.075 | 0.089** |
| | (0.108) | (0.053) | (0.060) | (0.061) | (0.057) | (0.041) |
| Y Mean in Target | – | 0.556 | 0.387 | 0.437 | 0.556 | |
| *B. Economic Sufficiency Index, Age 30 (PSID)* | | | | | | |
| Head Start | – | -0.023 | -0.045 | -0.025 | 0.025 | -0.002 |
| | – | (0.102) | (0.085) | (0.092) | (0.088) | (0.083) |
| Y Mean in Target | – | 0.213 | -0.198 | -0.485 | 0.213 | |
| *C. High School Graduation (CNLSY)* | | | | | | |
| Head Start | 0.086*** | 0.085*** | 0.043 | 0.051* | 0.024 | 0.035* |
| | (0.031) | (0.031) | (0.031) | (0.029) | (0.034) | (0.021) |
| Y Mean in Target | – | 0.776 | 0.734 | 0.766 | 0.776 | |
| *D. Idle (CNLSY)* | | | | | | |
| Head Start | -0.071* | -0.072* | -0.054 | -0.047 | -0.060 | 0.025 |
| | (0.038) | (0.038) | (0.038) | (0.037) | (0.041) | (0.025) |
| Y Mean in Target | – | 0.197 | 0.221 | 0.201 | 0.197 | |
| *E. Learning Disability (CNLSY)* | | | | | | |
| Head Start | -0.059*** | -0.059*** | -0.034* | -0.044** | -0.038** | 0.015 |
| | (0.021) | (0.021) | (0.019) | (0.018) | (0.019) | (0.014) |
| Y Mean in Target | – | 0.051 | 0.055 | 0.041 | 0.051 | |
| *F. Poor Health (CNLSY)* | | | | | | |
| Head Start | -0.070*** | -0.069*** | -0.056** | -0.066** | -0.049* | 0.003 |
| | (0.026) | (0.027) | (0.027) | (0.027) | (0.029) | (0.018) |
| Y Mean in Target | – | 0.103 | 0.098 | 0.074 | 0.103 | |

Notes: Column 1 of this table shows the FFE estimated impacts of Head Start for whites from GTC or for the whole sample from Deming (2009). Column 2 shows the FFE estimate using our expanded sample for PSID outcomes and using our replication sample for CNLSY outcomes. The outcomes in Panels A and B are taken from the PSID white sample, and the outcomes in Panels C to F are taken from the CNLSY sample. Columns 3 to 5 present reweighted estimates of the effect of Head Start for four target populations (shown in the column header) using the post-regression reweighting procedure described in the text. Column 6 presents the difference in the estimate in column 2 (FFE) and column 4 (reweighted for participants), with the standard error obtained from a bootstrap procedure described in the text. "–" is used to indicate that the information is not available. Sample size is N=2,986 for the expanded sample, and 1,036 for GTC. Standard errors are clustered on mother id. * $p < .10$, ** $p < .05$, *** $p < .01$.

# A Derivations and Proofs

## A.1 Proof of Proposition 1

The proof of Proposition 1 closely follows the proofs of Theorem 2 in Angrist and Fernandez-Val (2013) and Theorem 1 in Aronow and Carnegie (2013). There are two key differences. First, we rely on Group ID Conditional Independence (Assumption 1), instead of the IV exclusion restriction. Second, we condition on two propensity scores, unlike Aronow and Carnegie (2013), who condition on $Pr(D_i = 1)$, and Angrist and Fernandez-Val (2013), who condition on discrete covariates.

Recall that we define $\hat{\delta}^t := \frac{1}{\sum_i \mathbf{1}(g(i)\in G_S)} \sum_i \hat{w^t_{g(i)}} \cdot \hat{\delta}_g$ and $w^t_{g(i)} := \frac{Q_x}{P_x}\frac{P(S=1)}{P(T=1)}$.

By Assumptions 1 and 3,

$$\hat{\delta}^t \to_p \mathbb{E}\left[w^t_{g(i)} \cdot \hat{\delta}_g | S_g = 1\right] \tag{17}$$

By Assumption 1 and the law of iterated expectations,

$$\mathbb{E}\left[w^t_{g(i)} \cdot \hat{\delta} | S_g = 1\right] = \mathbb{E}\left[w^t_{g(i)} \left(Y_i\left(1\right) - Y_i\left(0\right)\right) | S_g = 1\right] \tag{18}$$

By the law of iterated expectations,

$$\begin{aligned}
\mathbb{E}[w^t_{g(i)} \cdot (Y_i\left(1\right) - Y_i\left(0\right)) | S = 1] &= \mathbb{E}[\mathbb{E}[w^t_{g(i)}(Y\left(1\right) - Y\left(0\right)) | S = 1, P_x, Q_x] | S = 1] && (19)\\
&= \mathbb{E}[\mathbb{E}[w^t_{g(i)}(Y\left(1\right) - Y\left(0\right)) | P_x, Q_x] | S = 1] && (20)\\
&= \mathbb{E}[w^t_{g(i)}\mathbb{E}[(Y(1) - Y(0)) | P_x, Q_x]) | S = 1] && (21)\\
&= \mathbb{E}[w^t_{g(i)}\Delta(P_x, Q_x) | S = 1]
\end{aligned}$$

Line (20) follows from (19) from the CFEI assumption that $\mathbb{E}[Y(1) - Y(0)|S, P_x, Q_x] = \mathbb{E}[Y(1) - Y(0)|P_x, Q_x]$. Line (21) follows from line (20) because $w$ is a function of $P_x$ and $Q_x$ only, by definition.

Now let $F$ be the distribution of $(P_x, Q_x)$, and let $F(\cdot|S = 1)$ be the distribution conditional on $S = 1$. By Bayes rule,

$$\begin{aligned}
\int w^t_g \Delta(P_x, Q_x) dF(P_x, Q_x | S = 1) &= \int w^t_{g(i)} \cdot \Delta(P_x, Q_x) \frac{Pr(S = 1|P_x, Q_x)}{Pr(S = 1)} dF(P_x, Q_x)\\
&= \int w^t_{g(i)} \cdot \Delta(P_x, Q_x) \frac{P_x}{Pr(S = 1)} dF(P_x, Q_x)\\
&= \int \frac{Q_x}{P_x}\frac{P(S = 1)}{P(T = 1)} \Delta(P_x, Q_x) \frac{P_x}{Pr(S = 1)} dF(P_x, Q_x)\\
&= \int \Delta(P_x, Q_x) \frac{Q_x}{Pr(T = 1)} dF(P_x, Q_x)\\
&= \int \Delta(P_x, Q_x) \frac{Pr(T = 1|P_x, Q_x)}{Pr(T = 1)} dF(P_x, Q_x)\\
&= \int \Delta(P_x, Q_x) dF(P_x, Q_x | T = 1)
\end{aligned}$$

By CFEI, $\int \Delta(P_x, Q_x) dF(P_x, Q_x|T=1) = \int \mathbb{E}[Y(1) - Y(0)|T=1, P_x, Q_x] dF(P_x, Q_x|T=1) = \mathbb{E}[Y(1) - Y(0)|T=1]$.

## A.2 Intuition for Propensity Score Weighting

In this section, we provide a simple derivation of the weighting scheme that we propose to obtain the ATE from the switchers sample by introducing a concrete example in which treatment effect is determined by one discrete covariate, X, and in which there are only few groups in the switcher sample. For ease of exposition, we refer to groups as families and units within groups as kids.

### A.2.1 Thought Experiment

Suppose that the target population is comprised of 75% black individuals and 25% white individuals. The switchers sample has 1 white family with 3 kids and 2 black families with 3 and 5 kids, respectively. Thus, to be representative of the target population, the white family should be given a weight of 25%. The share for each black family is proportional to the number of individuals in the family, normalized so that the total share across the two families is 75%. Thus, the first family should be given a weight of $0.75 \times \frac{3}{8}$, and the second family should be given a weight of $0.75 \times \frac{5}{8}$.

### A.2.2 Notation

Under the setup above, the weight that should be given to a switcher family $g$ where all individuals have race $x_i = x$, can be written as:

$$s_{gx} = \frac{\sum(\mathbf{1}(x_i = x)|T_{g(i)} = 1)}{\sum T_{g(i)}} \cdot \frac{\sum \mathbf{1}(g(i) = g)}{\sum(S_{g(i)}|x_i = x)} \tag{22}$$

The first term, $\frac{\sum(\mathbf{1}(x_i=x)|T_{g(i)}=1)}{\sum T_{g(i)}}$, gives the share of individuals in the target population with race x. The second term, $\frac{\sum \mathbf{1}(g(i)=g)}{\sum(S_{g(i)}|x_i=x)}$ gives the size of family g as a proportion of the switcher sample with race x.

Equivalently,

$$s_{gx} = pr(x_i = x|T_{g(i)} = 1) \cdot \frac{\sum \mathbf{1}(g(i) = g)}{pr(x_i = x|S_{g(i)} = 1) \times \sum S_{g(i)}} \tag{23}$$

$$= \frac{pr(x_i = x|T_{g(i)} = 1)}{pr(x_i = x|S_{g(i)} = 1)} \cdot pr(g(i) = g|S_{g(i)} = 1) \tag{24}$$

### A.2.3 Estimation

1. We obtain an estimate of $\hat{Q}_x = pr(T_{g(i)}|x_i = x)$ as fitted values from a regression of $T$ on $X$. This is equal to $\frac{pr(x_{g(i)}=x|T_{g(i)}=1) \cdot pr(T_{g(i)}=1)}{pr(x_i=x)}$ by Bayes rule.

2. We obtain an estimate of $\hat{P}_x = pr(S_{g(i)} = 1|x_i = x)$ as fitted values from a regression of $S$ on $X$.

This is equal to $\frac{pr(x_i=x|S_g=1)\cdot pr(S_{g(i)}=1)}{pr(x_i=x)}$ by Bayes rule. The ratio of (1) and (2) is $\frac{pr(x_i=x|T_{g(i)}=1)}{pr(x_i=x|S_{g(i)}=1)}\times \frac{pr(T_{g(i)}=1)}{pr(S_{g(i)}=1)}$.

3. To get $s_{gx}$, we need to multiply this ratio by $pr(g(i)=g|S_{g(i)}=1)$ and divide by $\frac{pr(T_{g(i)}=1)}{pr(S_{g(i)}=1)}$. We then normalize the weights, which gives $s_{gx}=\frac{\frac{Q_x}{P_x}\cdot n_g}{\sum_{g\in G_S}\frac{Q_x}{P_x}\cdot n_g}$

## A.3 Extension to Unit i Covariates

### A.3.1 Modified Assumptions

We begin with a simplification of the model, in which outcomes are a linear function of treatment, individual covariates, and additively separable individual error terms: $Y_{ig}=\delta_g\cdot D_i+\beta\cdot C_i+\alpha_g+(u_{ig}\cdot D_i+\epsilon_{ig})$. We assume constant coefficients on $C_i$, and require the systematic part of the treatment effect to be constant within a group. There can also be an idiosyncratic component of the treatment effect, denoted by $u_{ig}$.

Modified Assumption 1: $\epsilon_{ig}, u_{ig} \perp\!\!\!\perp D_i|C_i,\alpha_g$.

This gives that $E\left[\hat{\delta}_g\right]=\delta_g$.

We also now allow for individual covariates to enter into the propensity to be in the switching or target populations, respectively, as: $P_{X,C}=Pr\left[S_{g(i)}=1|X_g,C_i\right]$ and $Q_{X,C}=Pr\left[T_i=1|X_g,C_i\right]$.

The IPW weights therefore vary at the indivdiual level: $w_i=\frac{Q_{X,C}}{P_{X,C}}\cdot\frac{Pr(S=1)}{Pr(T-1)}$.

Assumptions 2-4 and Proposition 1 will carry forward with these redefined terms.

We can also relax the model to allow for covariate-dependent treatment effects: $Y_{ig}=\delta_g\cdot D_i+\beta_C\cdot(C_iD_i)+\beta\cdot C_i+\alpha_g+(u_{ig}\cdot D_i+\epsilon_{ig})$. To recover the ATE for a target group, first estimate the model allowing for an interaction effect, to recover $\hat{\delta}_g$ and $\hat{\beta}_C$. Then apply the second stage reweighting to the g-level components $\frac{1}{\sum_i \mathbf{1}(g(i)\in G_S)}\sum_{i|g(i)\in G_S} w_i\hat{\delta}_g$ to estimate $E\left[\delta_g|T_i=1\right]$. Finally, add this to $\hat{\beta}_C\cdot E\left[C_i|T_i=1\right]$ to obtain an estimated ATE for target individuals.

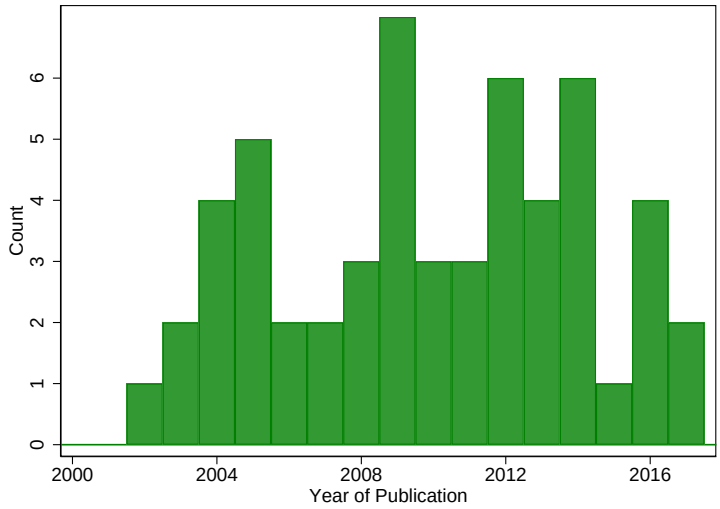
### A.3.2 Defining residual switchers

Consider the "deviations from group means" projection matrix, $M=I-H\left(H'H\right)^{-1}H$, with $H$ a matrix of dummy variables for group membership: $H\left[i,j\right]=1$ if unit $i$ is a member of group $j$, and 0 otherwise. Let $\widetilde{D}=M\cdot D$ be deviations in treatment from group means. Basic switcher groups ("true switchers") are defined by having within-group variation in treatment: $V_g:=Var\left(\widetilde{D}_i|g(i)=g\right)>0$. Next consider the residual-maker matrix projecting on covariates $C$ after taking deviations from group means, $L=I-(M\cdot C)\left(C'\cdot M\cdot C\right)(C'\cdot M)$, and let $\ddot{D}=L\cdot M\cdot D$. With unit-varying covariates $C_i$, the variation that identifies the treatment effects is $V_{g,C}:=Var\left(\ddot{D}|g(i)=g\right)$. For some groups $g$, it could be the case that $V_g=0$, and also $V_{g,C}>0$. We call these groups residual switchers.

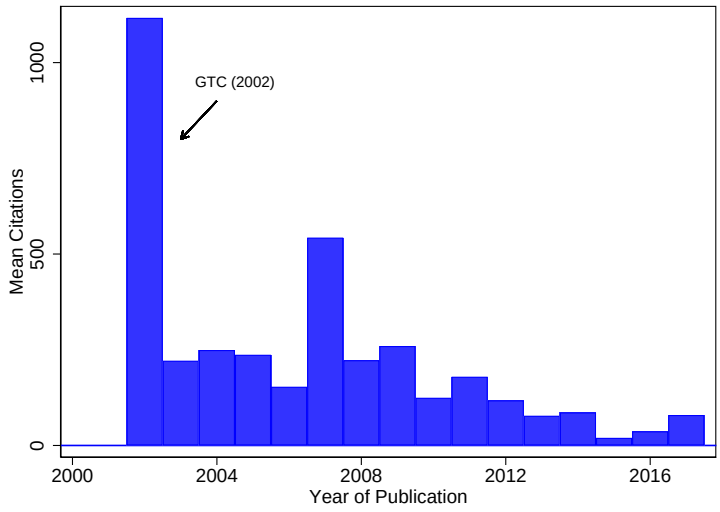
# B Supplementary Figures and Tables

Figure B.1: Popularity of Family Fixed Effects Articles

(a) Publications by Year



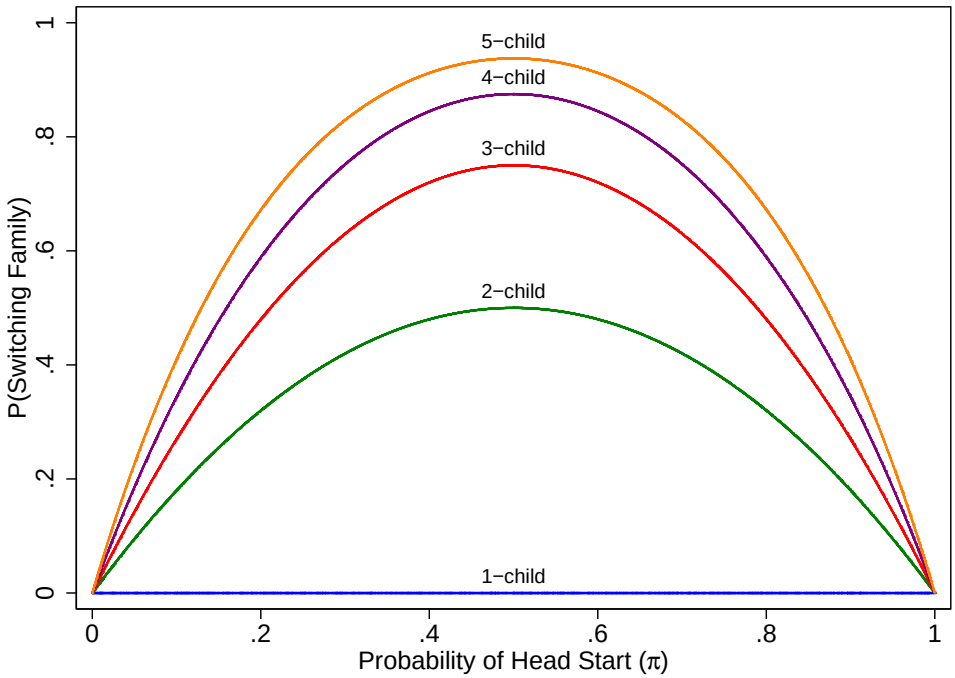

(b) Average Citations by Year of Publication




Notes: These figures display the data from our survey of FFE papers published from January 2000 to May 2017 in 11 leading journals that publish applied microeconomics articles. Figure (a) plots the number of FFE articles published in each year, and Figure (b) plots the average number of Google Scholar citations, as of May 2019, among the articles published in a given year.

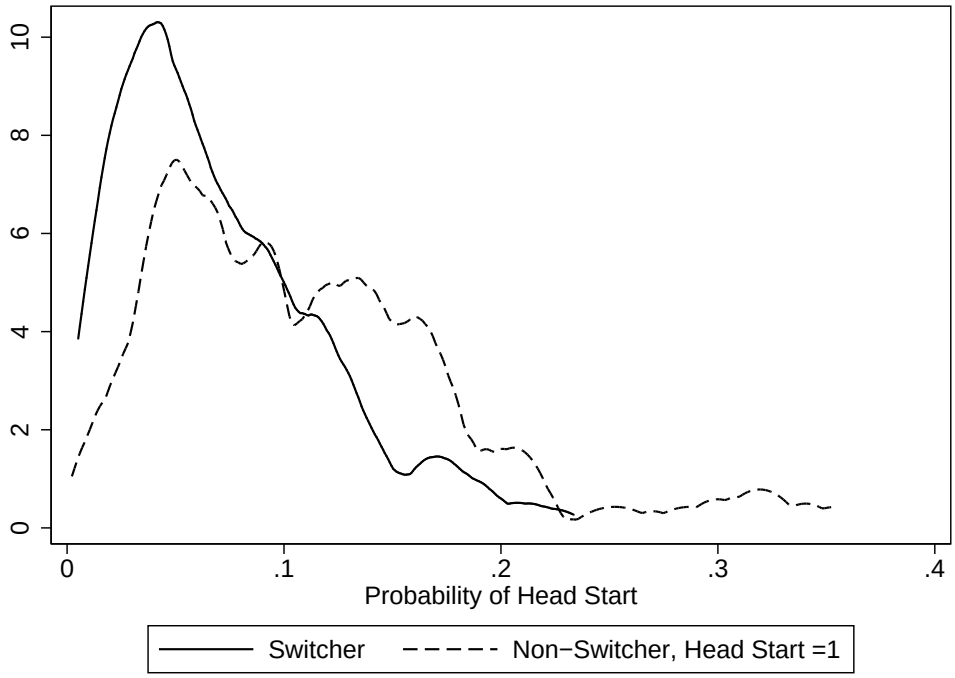
Figure B.2: Illustrative Model of the Role of Family Size in Switching

$$P(HSSwitchingFamily) = 1 - (1 - \pi)^{n_g} - \pi^{n_g}$$



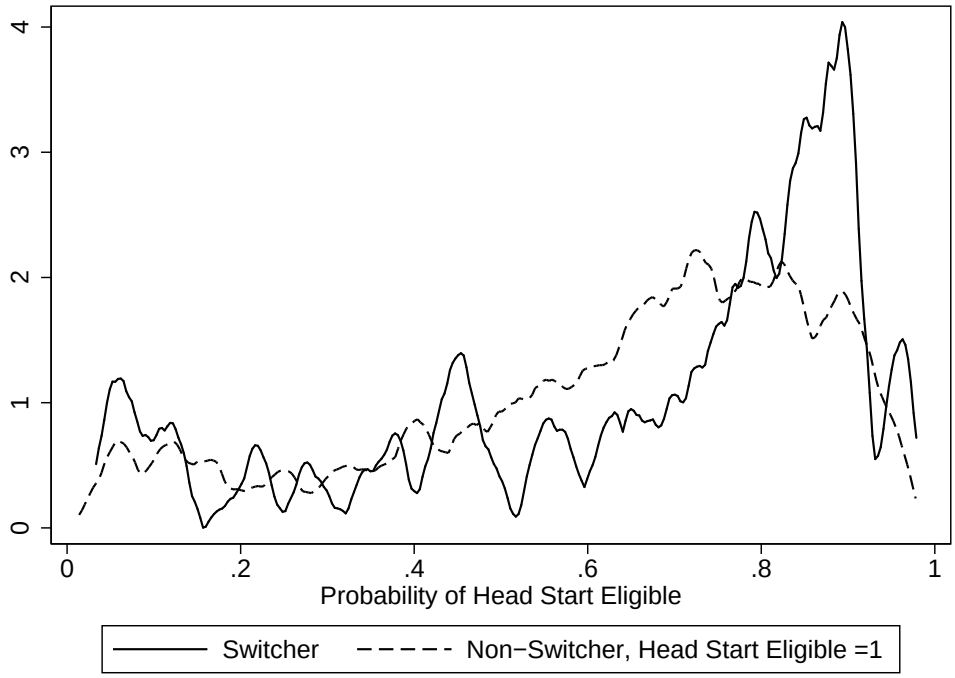

Notes: This figure plots the theoretical function: $P(HSSwitchingFamily) = 1 - (1 - \pi)^{n_g} - \pi^{n_g}$, where $n_g$ is the number of children in a family and $\pi$ is the probability of attending Head Start, for 2-, 3-, 4-, and 5 (plus)- child families.

Figure B.3: Examining P-Score Overlap: Predicted Probability of Being in Head Start (PSID White Sample)




Notes: This figure shows kernel density plots (bandwidth = 0.01) of the predicted probability of being a Head Start participant for switchers and non-switchers that are Head Start participants. The sample consists of white individuals in the PSID.

Figure B.4: Examining P-Score Overlap: Predicted Probability of Being Head-Start-Eligible (PSID White Sample)




Notes: This figure shows kernel density plots (bandwidth = 0.01) of the predicted probability of being Head Start eligible for switchers and non-switchers that are Head-Start-eligible. The sample consists of white individuals in the PSID.

Table B.1: Head Start Attendance and Within-Family Variation in Attendance by Family Size (PSID)

| | Number of Children in Family: | | | | |
|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5+ | Total |
| Share of Family in Head Start ($\pi$) | 0.157 | 0.222 | 0.195 | 0.206 | 0.182 |
| Share with Switching | 0.121 | 0.202 | 0.242 | 0.471 | 0.174 |
| All Participants in HS in Family | 0.096 | 0.125 | 0.093 | 0.049 | 0.102 |
| No Participants in HS in Family | 0.783 | 0.672 | 0.665 | 0.480 | 0.724 |

Notes: This table shows the sources of switching by family size. The first two rows show the likelihood of attending Head Start by family size and the likelihood of having variation in Head Start within a family (switching). The final two rows examines whether differences in rates of switching across family sizes are attributable to variation across family sizes in having all children attend Head Start (row 3) or variation in having no children attend Head Start (row 4).

Table B.2: Demographic Characteristics of Head Start Sample (PSID)

| | All | Head Start | No Head Start | Sibling Sample |
|---|---|---|---|---|
| Head Start | 0.076 | 1.000 | 0.000 | 0.073 |
| Other preschool | 0.282 | 0.000 | 0.305 | 0.259 |
| Fraction African-American | 0.150 | 0.618 | 0.111 | 0.154 |
| Fraction female | 0.504 | 0.548 | 0.501 | 0.501 |
| Fraction low birth weight | 0.060 | 0.114 | 0.056 | 0.061 |
| Had a single mother at age 4 | 0.112 | 0.296 | 0.091 | 0.103 |
| Fraction whose mother completed hs | 0.717 | 0.632 | 0.724 | 0.689 |
| Fraction whose father completed hs | 0.683 | 0.557 | 0.692 | 0.654 |
| Fraction eldest child in family | 0.368 | 0.341 | 0.371 | 0.339 |
| Age in 1995 | 23.830 | 18.605 | 24.262 | 25.063 |
| | (9.84) | (7.76) | (9.87) | (10.06) |
| Mother's yrs education | 11.116 | 10.208 | 11.190 | 10.942 |
| | (2.76) | (2.32) | (2.78) | (2.81) |
| Father's yrs education | 11.238 | 10.159 | 11.314 | 11.076 |
| | (3.23) | (2.70) | (3.25) | (3.35) |
| Family income (age 3-6) (CPI adjusted) | 50339 | 28553 | 52719 | 50973 |
| | (35814.01) | (17212.32) | (36509.36) | (37315.99) |
| Household size at age 4 | 4.535 | 4.814 | 4.504 | 4.778 |
| | (1.68) | (2.06) | (1.63) | (1.64) |
| Observations | 7363 | 1345 | 6018 | 5355 |

Notes: This table shows the mean demographic characteristics of the sample, weighted to be representative of 1995 population; see text for details. Standard deviations, shown in parentheses, are omitted for binary variables. CPI-adjusted income reported in 1999 dollars. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table B.3: Outcomes of Interest for Head Start Sample (PSID)

| | All | Head Start | No Head Start | Sibling Sample |
|---|---|---|---|---|
| Fraction completed hs | 0.913 | 0.878 | 0.916 | 0.912 |
| Fraction attended some college | 0.531 | 0.428 | 0.539 | 0.532 |
| Fraction not booked/charged with crime | 0.899 | 0.889 | 0.900 | 0.898 |
| Avg. Earnings age 23-25 (CPI adjusted) | 20410 | 14391 | 20818 | 20633 |
| | (24927) | (12000) | (25517) | (26547) |
| Economic Sufficiency Index at 30 | 0.094 | -0.601 | 0.151 | 0.096 |
| | (1.03) | (1.05) | (1.01) | (1.03) |
| Economic Sufficiency Index at 40 | 0.020 | -0.532 | 0.053 | 0.025 |
| | (1.01) | (0.95) | (1.01) | (1.04) |
| Good Health Index at 30 | 0.004 | -0.558 | 0.050 | 0.017 |
| | (1.03) | (1.26) | (0.99) | (0.99) |
| Good Health Index at 40 | 0.011 | -0.486 | 0.033 | 0.015 |
| | (1.01) | (1.25) | (1.00) | (0.96) |
| Observations | 7363 | 1345 | 6018 | 5355 |

Notes: This table shows the means for the main outcomes of interest, weighted to be representative of 1995 population; see text for details. Note that the fraction not booked/charged with a crime restricted to individuals that responded to the PSID in 1995 who were between the ages of 16 and 50 in that year. CPI-adjusted income reported in 1999 dollars. Standard deviations, shown in parentheses, are omitted for binary variables. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table B.4: Summary Statistics for Inputs to Summary Indices (PSID)

| | All | Head Start | No Head Start | Sibling Sample |
|---|---|---|---|---|
| *Inputs to Economic Sufficiency Index, 30* | | | | |
| Ever on AFDC/TANF by age 30 | 0.062 | 0.220 | 0.049 | 0.060 |
| Fraction of last 5 yrs on Food Stamps/SNAP, age 30 | 0.064 | 0.151 | 0.056 | 0.071 |
| | (0.20) | (0.30) | (0.19) | (0.22) |
| ln(mean earnings in last 5 years), age 30 | 9.661 | 9.415 | 9.676 | 9.659 |
| | (1.06) | (0.91) | (1.07) | (1.07) |
| Fraction of last 5 yrs with positive earnings, age 30 | 0.895 | 0.887 | 0.896 | 0.898 |
| | (0.25) | (0.26) | (0.25) | (0.25) |
| Fraction of last 5 yrs ever unemployed, age 30 | 0.146 | 0.173 | 0.144 | 0.150 |
| | (0.24) | (0.27) | (0.23) | (0.24) |
| Mean Inc. Rel. Pov. in last 5 years, age 30 | 385.831 | 233.796 | 396.729 | 385.933 |
| | (305.98) | (155.44) | (311.18) | (291.36) |
| Fraction completed college | 0.209 | 0.073 | 0.220 | 0.220 |
| *Inputs to Economic Sufficiency Index, 40* | | | | |
| Ever on AFDC/TANF by age 40 | 0.068 | 0.163 | 0.062 | 0.067 |
| Fraction of last 5 yrs on Food Stamps/SNAP, age 40 | 0.043 | 0.098 | 0.040 | 0.043 |
| | (0.16) | (0.25) | (0.16) | (0.16) |
| ln(mean earnings in last 5 years), age 40 | 9.962 | 9.779 | 9.968 | 9.957 |
| | (1.15) | (0.90) | (1.16) | (1.15) |
| Fraction of last 5 yrs with positive earnings, age 40 | 0.850 | 0.867 | 0.849 | 0.849 |
| | (0.31) | (0.29) | (0.31) | (0.31) |
| Fraction of last 5 yrs ever unemployed, age 40 | 0.094 | 0.122 | 0.093 | 0.098 |
| | (0.20) | (0.24) | (0.19) | (0.20) |
| Mean Inc. Rel. Pov. in last 5 years, age 40 | 436.769 | 281.489 | 443.338 | 434.280 |
| | (366.03) | (183.89) | (370.36) | (361.58) |
| Fraction of last 5 yrs owned home, age 40 | 0.500 | 0.287 | 0.510 | 0.522 |
| | (0.44) | (0.42) | (0.44) | (0.44) |
| *Inputs to Good Health Index, 30* | | | | |
| Fraction of last 5 yrs smoked less than 1 cigarette/day, age 30 | 0.745 | 0.668 | 0.753 | 0.755 |
| | (0.41) | (0.45) | (0.41) | (0.40) |
| Fraction of last 5 yrs reported good or better health, age 30 | 0.948 | 0.903 | 0.951 | 0.950 |
| | (0.17) | (0.24) | (0.17) | (0.17) |
| Mean BMI in last 5 years, age 30 | 26.569 | 28.766 | 26.333 | 26.615 |
| | (6.68) | (6.74) | (6.63) | (6.85) |
| *Inputs to Good Health Index, 40* | | | | |
| Fraction of last 5 yrs smoked less than 1 cigarette/day, age 40 | 0.738 | 0.714 | 0.739 | 0.728 |
| | (0.42) | (0.44) | (0.42) | (0.42) |
| Fraction of last 5 yrs reported good or better health, age 40 | 0.919 | 0.871 | 0.921 | 0.922 |
| | (0.22) | (0.29) | (0.22) | (0.22) |
| Mean BMI in last 5 years, age 40 | 27.504 | 30.191 | 27.327 | 27.433 |
| | (5.92) | (7.42) | (5.77) | (5.85) |
| Observations | 7363 | 1345 | 6018 | 5355 |

Notes: Weighted to be representative of 1995 population; see text for details. SD, in parentheses, are omitted for binary variables. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table B.5: N's for Control Covariates (PSID)

| | All | Head Start | No Head Start | Sibling Sample |
|---|---|---|---|---|
| Head Start | 7372 | 1354 | 6018 | 5361 |
| Other preschool | 7372 | 1354 | 6018 | 5361 |
| Fraction African-American | 7372 | 1354 | 6018 | 5361 |
| Fraction female | 7372 | 1354 | 6018 | 5361 |
| Fraction low birth weight | 5366 | 970 | 4396 | 4555 |
| Had a single mother at age 4 | 6678 | 1285 | 5393 | 4672 |
| Fraction whose mother completed hs | 7231 | 1332 | 5899 | 5360 |
| Fraction whose father completed hs | 6596 | 1034 | 5562 | 4875 |
| Fraction eldest child in family | 7372 | 1354 | 6018 | 5361 |
| Age in 1995 | 7372 | 1354 | 6018 | 5361 |
| Mother's yrs education | 7223 | 1331 | 5892 | 5356 |
| Father's yrs education | 6596 | 1034 | 5562 | 4875 |
| Family income (age 3-6) (CPI adjusted) | 6086 | 1145 | 4941 | 4338 |
| Household size at age 4 | 6251 | 1187 | 5064 | 4420 |
| Observations | 7372 | 1354 | 6018 | 5361 |

Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table B.6: N's for Main Outcomes (PSID)

| | All | Head Start | No Head Start | Sibling Sample |
|---|---|---|---|---|
| Fraction completed hs | 7372 | 1354 | 6018 | 5361 |
| Fraction attended some college | 7372 | 1354 | 6018 | 5361 |
| Fraction not booked/charged with crime | 5005 | 802 | 4203 | 3591 |
| Avg. Earnings age 23-25 (CPI adjusted) | 4866 | 783 | 4083 | 3675 |
| Economic Sufficiency Index at 30 | 7372 | 1354 | 6018 | 5361 |
| Economic Sufficiency Index at 40 | 4085 | 613 | 3472 | 2845 |
| Good Health Index at 30 | 4749 | 791 | 3958 | 3600 |
| Good Health Index at 40 | 2228 | 312 | 1916 | 1673 |
| Observations | 7372 | 1354 | 6018 | 5361 |

Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table B.7: N's for Auxiliary Outcomes (PSID)

| | All | Head Start | No Head Start | Sibling Sample |
|---|---|---|---|---|
| *Inputs to Economic Sufficiency Index, 30* | | | | |
| Ever on AFDC/TANF by age 30 | 7372 | 1354 | 6018 | 5361 |
| Fraction of last 5 yrs on Food Stamps/SNAP, age 30 | 4186 | 713 | 3473 | 2805 |
| ln(mean earnings in last 5 years), age 30 | 4202 | 620 | 3582 | 3159 |
| Fraction of last 5 yrs with positive earnings, age 30 | 4378 | 656 | 3722 | 3295 |
| Fraction of last 5 yrs ever unemployed, age 30 | 4259 | 634 | 3625 | 3184 |
| Mean Inc. Rel. Pov. in last 5 years, age 30 | 5293 | 891 | 4402 | 4068 |
| Fraction completed college | 7372 | 1354 | 6018 | 5361 |
| *Inputs to Economic Sufficiency Index, 40* | | | | |
| Ever on AFDC/TANF by age 40 | 4085 | 613 | 3472 | 2845 |
| Fraction of last 5 yrs on Food Stamps/SNAP, age 40 | 1972 | 250 | 1722 | 1423 |
| ln(mean earnings in last 5 years), age 40 | 1695 | 221 | 1474 | 1266 |
| Fraction of last 5 yrs with positive earnings, age 40 | 1829 | 236 | 1593 | 1369 |
| Fraction of last 5 yrs ever unemployed, age 40 | 1825 | 236 | 1589 | 1365 |
| Mean Inc. Rel. Pov. in last 5 years, age 40 | 2152 | 296 | 1856 | 1613 |
| Fraction of last 5 yrs owned home, age 40 | 2292 | 290 | 2002 | 1625 |
| *Inputs to Good Health Index, 30* | | | | |
| Fraction of last 5 yrs smoked less than 1 cigarette/day, age 30 | 2267 | 385 | 1882 | 1742 |
| Fraction of last 5 yrs reported good or better health, age 30 | 3763 | 579 | 3184 | 2806 |
| Mean BMI in last 5 years, age 30 | 3248 | 587 | 2661 | 2528 |
| *Inputs to Good Health Index, 40* | | | | |
| Fraction of last 5 yrs smoked less than 1 cigarette/day, age 40 | 1280 | 182 | 1098 | 930 |
| Fraction of last 5 yrs reported good or better health, age 40 | 1463 | 182 | 1281 | 1116 |
| Mean BMI in last 5 years, age 40 | 2037 | 307 | 1730 | 1486 |
| Observations | 7372 | 1354 | 6018 | 5361 |

Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table B.8: Effect of Head Start on Pre-Head-Start Outcomes (PSID)

| | All | Sibs | Mom FE | Blk, FE | Wht, FE |
|---|---|---|---|---|---|
| *Low birth weight* | | | | | |
| Head Start | 0.040* | 0.045* | -0.016 | -0.018 | -0.029 |
| | (0.021) | (0.023) | (0.026) | (0.033) | (0.042) |
| Other preschool | 0.003 | 0.003 | -0.012 | -0.056** | -0.003 |
| | (0.012) | (0.013) | (0.023) | (0.027) | (0.027) |
| Observations | 5366 | 4555 | 4500 | 1872 | 2622 |
| *Disabled* | | | | | |
| Head Start | -0.006 | -0.017 | -0.010 | -0.016 | -0.006 |
| | (0.027) | (0.030) | (0.030) | (0.036) | (0.051) |
| Other preschool | 0.018 | 0.018 | 0.021 | 0.032 | 0.017 |
| | (0.019) | (0.022) | (0.028) | (0.049) | (0.032) |
| Observations | 3516 | 2955 | 2661 | 1102 | 1555 |
| *Single mom at age 4* | | | | | |
| Head Start | 0.020 | 0.025 | 0.027 | -0.007 | 0.051 |
| | (0.015) | (0.020) | (0.024) | (0.022) | (0.040) |
| Other preschool | 0.022** | 0.020* | 0.008 | 0.006 | 0.011 |
| | (0.009) | (0.011) | (0.017) | (0.031) | (0.018) |
| Observations | 6678 | 4672 | 4467 | 1939 | 2522 |
| *Family income (age 1) (CPI adjusted)* | | | | | |
| Head Start | 0.000** | -0.000*** | 0.000 | -0.000 | -0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Other preschool | -0.000*** | -0.000*** | -0.000 | 0.000 | -0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Observations | 6219 | 4313 | 4023 | 1719 | 2298 |
| *Family income (age 2) (CPI adjusted)* | | | | | |
| Head Start | 0.000 | -0.000 | -0.000 | 0.000 | -0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Other preschool | -0.000 | -0.000 | -0.000 | -0.000 | -0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Observations | 6274 | 4391 | 4151 | 1757 | 2388 |
| *Mom working at age 1* | | | | | |
| Head Start | 0.001 | 0.011 | 0.049 | 0.002 | 0.080 |
| | (0.018) | (0.022) | (0.039) | (0.033) | (0.073) |
| Other preschool | -0.001 | -0.002 | -0.017 | -0.078* | -0.014 |
| | (0.013) | (0.016) | (0.030) | (0.043) | (0.034) |
| Observations | 6219 | 4313 | 4023 | 1719 | 2298 |
| *Mom working at age 2* | | | | | |
| Head Start | 0.025 | 0.028 | -0.041 | -0.008 | -0.077 |
| | (0.021) | (0.023) | (0.040) | (0.036) | (0.073) |
| Other preschool | 0.026* | 0.032* | 0.015 | -0.013 | 0.017 |
| | (0.015) | (0.018) | (0.031) | (0.044) | (0.036) |
| Observations | 6274 | 4391 | 4151 | 1757 | 2388 |

Notes: Weighted to be representative of 1995 population; see text for details. SE clustered at 1968 family id in columns 1 and 2 and at mother id level otherwise. * p < .10, ** p < .05, *** p < .01. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table B.9: Test of Conditional Independence Assumption:

Do Individuals in the Target Population Have Differential Treatment Effects?

| | Eligible | Participants |
|---|---|---|
| *Some College (Whites, PSID)* | | |
| In Target | -0.056 | -0.072 |
| | (0.058) | (0.058) |
| Observations | 306 | 315 |
| *Economic Sufficiency Index (Whites, PSID)* | | |
| In Target | -0.006 | -0.054 |
| | (0.084) | (0.089) |
| Observations | 306 | 315 |
| *High School Graduation (CNLSY)* | | |
| In Target | 0.006 | 0.005 |
| | (0.030) | (0.019) |
| Observations | 1012 | 1251 |
| *Idle (CNLSY)* | | |
| In Target | 0.015 | -0.017 |
| | (0.038) | (0.024) |
| Observations | 1012 | 1251 |
| *Learning Disability (CNLSY)* | | |
| In Target | -0.030 | -0.024* |
| | (0.019) | (0.013) |
| Observations | 1012 | 1251 |
| *Poor Health (CNLSY)* | | |
| In Target | -0.009 | -0.035* |
| | (0.027) | (0.018) |
| Observations | 1012 | 1251 |

Notes: Each cell of this table shows an estimate from a regression of the family-specific impact of Head Start on an indicator for whether an individual is in the target population. Regressions are weighted by our constructed propensity score weights. The first two panels use data from the PSID white sample, and the final four panels use data from the CNLSY.

Table B.10: Additional Estimates for Representative White Populations (PSID)

Using Post-Regression Reweighting Method

| | FFE | | Reweighted, Target = | | |
|---|---|---|---|---|---|
| | GTC | Expand Sample | HS Eligible | Participants | Siblings |
| *A. High School Graduation* | | | | | |
| Head Start | 0.203** | -0.015 | -0.036 | -0.033 | -0.030 |
| | (0.098) | (0.045) | (0.043) | (0.047) | (0.051) |
| Y Mean in Target | – | 0.921 | 0.852 | 0.848 | 0.921 |
| *B. Good Health Index, Age 30* | | | | | |
| Head Start | – | -0.265 | -0.226 | -0.423 | -0.157 |
| | – | (0.249) | (0.267) | (0.307) | (0.319) |
| Y Mean in Target | – | 0.074 | -0.061 | -0.583 | 0.074 |

Notes: Columns 1 and 2 of this table show the FFE estimated impacts of Head Start from GTC (2002) and using our expanded sample for completion of high school (panel A) and the Good Health Index at age 30 (panel B). The remaining columns present reweighted estimates of the effect of Head Start for three target populations (shown in the column header) using the post-regression reweighting procedure described in the text. "–" is used to indicate that the information is not available. Sample size is N=2,986 for the expanded sample in panel A, and 1,959 for the expanded sample in panel B, and 1,036 for GTC. Standard errors are clustered on mother id. * $p < .10$, ** $p < .05$, *** $p < .01$. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table B.11: Head Start Impact for Representative Black Eligible, Participants, and Siblings (PSID)

Using Post-Regression Reweighting Method

| | FFE | | Reweighted, Target = | | |
|---|---|---|---|---|---|
| | GTC | Expand Sample | HS Eligible | Participants | Siblings |
| *A. High School Graduation* | | | | | |
| Head Start | -0.025 | -0.024 | -0.018 | -0.015 | -0.016 |
| | (0.065) | ( 0.031) | (0.025) | (0.026) | ( 0.023) |
| Y Mean in Target | – | 0.862 | 0.854 | 0.896 | 0.862 |
| *B. Some College* | | | | | |
| Head Start | 0.023 | -0.016 | -0.029 | -0.029 | -0.029 |
| | (0.066) | (0.036) | (0.031) | (0.034) | (0.031) |
| Y Mean in Target | – | 0.396 | 0.376 | 0.423 | 0.396 |
| *C. Economic Sufficiency Index, Age 30* | | | | | |
| Head Start | – | -0.117 | -0.182*** | -0.208*** | -0.160** |
| | – | (0.081) | (0.071) | (0.072) | (0.070) |
| Y Mean in Target | – | -0.552 | -0.626 | -0.674 | -0.552 |
| *D. Good Health Index, Age 30* | | | | | |
| Head Start | – | 0.024 | 0.046 | 0.055 | 0.031 |
| | – | (0.149) | (0.145) | (0.161) | (0.134) |
| Y Mean in Target | – | -0.357 | -0.381 | -0.539 | -0.357 |

Notes: Columns 1 and 2 of this table show the FFE estimated impacts of Head Start from GTC (2002) and using our expanded sample for completion of high school (panel A) and the Good Health Index at age 30 (panel B). The remaining columns present reweighted estimates of the effect of Head Start for three target populations (shown in the column header) using the post-regression reweighting procedure described in the text. "–" is used to indicate that the information is not available. Sample size is N=2,369 for the expanded sample in panels A, B, and C, and 1,150 for the expanded sample in Panel D, and 762 for GTC. Standard errors are clustered on mother id. * $p < .10$, ** $p < .05$, *** $p < .01$. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table B.12: FFE Estimates Reweighted using Gibbons, Suarez, Urbancic (2018) Method Compared with FFE and Propensity Score Reweighting

| | FFE | GSU (2018) Reweight | Reweighted, Target = |
|---|---|---|---|
| | Baseline | Switchers | Siblings |
| *A. Some College (PSID)* | | | |
| Head Start | 0.120** | 0.134*** | 0.075 |
| | (0.053) | (0.053) | (0.057) |
| *B. Economic Sufficiency Index, Age 30 (PSID)* | | | |
| Head Start | -0.023 | -0.081 | 0.025 |
| | (0.102) | (0.094) | (0.088) |
| *C. High School Graduation (CNLSY)* | | | |
| Head Start | 0.085*** | 0.084*** | 0.024 |
| | (0.030) | (0.027) | (0.034) |
| *D. Idle (CNLSY)* | | | |
| Head Start | -0.072* | -0.068** | -0.060 |
| | (0.037) | (0.034 | (0.041) |
| *E. Learning Disability (CNLSY)* | | | |
| Head Start | -0.059*** | -0.053*** | -0.038** |
| | (0.020) | (0.019) | (0.019) |
| *F. Poor Health (CNLSY)* | | | |
| Head Start | -0.069*** | -0.059** | -0.049* |
| | (0.026) | (0.025) | (0.029) |

Notes: Column 1 reprints the FFE estimate using our expanded sample for PSID outcomes and using our replication sample for CNLSY outcomes. The outcomes in Panels A and B are taken from the PSID white sample, and the outcomes in Panels C to F are taken from the CNLSY sample. Column 2 presents the estimate weighting family-level estimates by the sample share, as suggested in Gibbons, Urbancic, Suarez Serrato (2018). This "undoes" the conditional variance weighting of FFE, and produces an estimate that is interpretable as the ATE for switchers. Column 3 reprints reweighted estimates of the effect of Head Start for siblings using the post-regression reweighting procedure described in the text. Sample size is N=2,986 for the expanded sample. Standard errors are clustered on mother id. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table B.13: Horse Race between Family Size and Index of Non-Family-Size Covariates (PSID White Sample)

| | x Fam Size | x Index | Horse Race |
|---|---|---|---|
| *Index = Predicted Head Start* | | | |
| Head Start | 0.025 | 0.073 | 0.008 |
| | (0.063) | (0.069) | (0.072) |
| Head Start x 4plus child family | 0.281** | | 0.250** |
| | (0.112) | | (0.105) |
| Head Start x Tercile 1 Predicted Head Start | | -0.049 | -0.116 |
| | | (0.094) | (0.101) |
| Head Start x Tercile 2 Predicted Head Start | | 0.212* | 0.125 |
| | | (0.113) | (0.111) |
| Observations | 2986 | 2986 | 2986 |
| *Index = Predicted Finish College* | | | |
| Head Start | 0.025 | -0.088 | -0.130 |
| | (0.063) | (0.083) | (0.100) |
| Head Start x 4plus child family | 0.281** | | 0.266** |
| | (0.112) | | (0.112) |
| Head Start x Tercile 1 Predicted Finish College | | 0.237** | 0.155 |
| | | (0.112) | (0.121) |
| Head Start x Tercile 2 Predicted Finish College | | 0.260** | 0.207 |
| | | (0.131) | (0.142) |
| Observations | 2986 | 2986 | 2986 |

This table shows estimates from a FFE regression of attainment of some college on an indicator for attendance of Head Start, and an indicator for having a family with 4 or more children (Column 1), dummies for terciles of an index of predicted Head Start attendance (Column 2, Panel A), dummies for terciles of an index of the predicted likelihood of finishing college (Column 2, Panel B), and the combination of family size indicator and terciles of the index (Column 3). The predicted Head Start (finish college) index is created by regressing Head Start attendance (finish college) on all of the control variables in the PSID analysis, except for the household size variable.

# C Supplementary PSID FFE Results

In this section, we discuss additional FFE results obtained using our expanded PSID sample.

We present the FFE results for the economic and health indices measured at age 40, together with the indices at age 30 for comparison, in Table C.1. Overall, the results suggest little support for a positive long term effect of Head Start. We come to the same conclusions when we aggregate the inputs using principal components analysis (see Table C.2. Our overall conclusions are not changed importantly by looking at specific outcomes or subsamples. We have also estimated regressions for each of the inputs to the economic and health indices, which we include in Tables C.3, and C.4, C.5, C.6. Table C.7 shows the regression results for the additional outcomes analyzed in GTC, earnings between ages 23 to 25, and not having committed a crime. Across these tables, there is no systematic evidence that Head Start impacts long term outcomes.[49]

Motivated by the prior findings of differential effects by gender in Carneiro and Ginja (2014); Deming (2009), in Table C.8 we look to see whether our mean results are obscuring this form of heterogeneity in our setting. Curiously, we find some evidence of significant negative effects of Head Start among men, in particular for health and economic outcomes at age 40. On the other hand, we find a positive and significant effect of Head Start on the probability that men attain some college. The effects estimated for women are never individually significant, but also not statistically different from men for many outcomes as indicated by the p-value of the difference in the table. The one exception is for economic outcomes observed at age 40, where women are found to have significantly better returns to Head Start participation than observed for men.

Another source of heterogeneity which could generate a discrepancy between our results and GTC is the fact that our sample includes later (younger) cohorts, whose Head Start experience may differ from earlier participants. In Table C.9, we find some support for a decreasing impact of Head Start across cohorts for the age 40 indices, but also find a larger improvement in the health index at age 30 for more recent cohorts. Thus, this does not appear to reconcile our findings.[50]

[49] Moreover, while we find a significant increase in ataiment of some college, when we examine the outcome of college completion, we obtain insignificant negative point estimates for the pooled sample (beta = -0.033, se = 0.023), for black children (beta = -0.014, se = 0.018), and for white children (beta = -0.058, se = 0.043).

[50] Moreover, when we instead use a binary indicator for more recent cohorts, we do not find a statistically significant difference in the impacts of Head Start, indicating that these results are sensitive to functional form assumptions.

Table C.1: Impact of Head Start on Economic Sufficiency Index and Good Health Index (PSID)

| | All | Sibs | Mom FE | Blk, FE | Wht, FE |
|---|---|---|---|---|---|
| *Economic Sufficiency Index, age 30* | | | | | |
| Head Start | -0.147*** | -0.117** | -0.090 | -0.117 | -0.023 |
| | (0.043) | (0.050) | (0.064) | (0.081) | (0.102) |
| Other preschool | 0.184*** | 0.181*** | 0.091 | 0.050 | 0.099 |
| | (0.035) | (0.040) | (0.062) | (0.109) | (0.072) |
| Mean Y | 0.094 | 0.096 | 0.096 | -0.552 | 0.213 |
| Observations | 7372 | 5361 | 5361 | 2369 | 2986 |
| *Economic Sufficiency Index, age 40* | | | | | |
| Head Start | -0.080 | -0.071 | -0.059 | -0.170 | -0.081 |
| | (0.066) | (0.077) | (0.100) | (0.134) | (0.125) |
| Other preschool | 0.112* | 0.085 | 0.043 | -0.270 | 0.118 |
| | (0.059) | (0.077) | (0.107) | (0.223) | (0.122) |
| Mean Y | 0.020 | 0.025 | 0.025 | -0.670 | 0.142 |
| Observations | 4085 | 2845 | 2503 | 1065 | 1435 |
| *Good Health Index, Age 30* | | | | | |
| Head Start | -0.349*** | -0.320*** | -0.148 | 0.024 | -0.265 |
| | (0.058) | (0.064) | (0.143) | (0.149) | (0.249) |
| Other preschool | 0.087** | 0.096** | 0.081 | 0.040 | 0.106 |
| | (0.038) | (0.045) | (0.076) | (0.159) | (0.084) |
| Mean Y | 0.004 | 0.017 | 0.017 | -0.357 | 0.074 |
| Observations | 4749 | 3600 | 3114 | 1150 | 1959 |
| *Good Health Index, Age 40* | | | | | |
| Head Start | -0.201* | -0.175 | -0.147 | 0.031 | -0.146 |
| | (0.118) | (0.141) | (0.202) | (0.201) | (0.393) |
| Other preschool | 0.117 | 0.095 | 0.119 | 0.382* | 0.038 |
| | (0.094) | (0.115) | (0.130) | (0.210) | (0.150) |
| Mean Y | 0.011 | 0.015 | 0.015 | -0.290 | 0.062 |
| Observations | 2228 | 1673 | 1306 | 511 | 795 |

Notes: This table shows the estimates from regressions of either the Economic Sufficiency Index at age 30 (panel A), the Economic Sufficiency Index at age 40 (panel B), the Good Health Index at age 30 (panel C), or the Good Health Index at age 40 (panel D) on an indicator for participation in Head Start and control variables described in the text. Regressions are run on the whole sample (column 1), siblings (columns 2 and 3), black siblings (column 4) and white siblings (column 5). All columns include control variables, and columns 3, 4, and 5 include mother fixed effects. The Good Health Index includes measures of not smoking cigarettes, good self reported health and BMI, averaged over the previous 5 years. The Economic Sufficiency Index includes measures of high school graduation, attendance of some college, no receipt of Food Stamps/SNAP, no receipt of AFDC/TANF, average earnings, employment, and unemployment, averaged over the previous 5 years. Estimates are weighted to be representative of 1995 population; see text for details. Standard errors are clustered at 1968 family id in column 1 and at mother id level otherwise. * p < .10, ** p < .05, *** p < .01. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table C.2: Effect of Head Start on Economic and Health Principal Components (PSID)

| | All | Sibs | Mom FE | Blk, FE | Wht, FE |
|---|---|---|---|---|---|
| *Economic Sufficiency Principal Component, age 30* | | | | | |
| Head Start | -0.174*** | -0.140** | -0.100 | -0.138 | -0.031 |
| | (0.058) | (0.068) | (0.084) | (0.109) | (0.128) |
| Other preschool | 0.295*** | 0.285*** | 0.150* | 0.071 | 0.166 |
| | (0.051) | (0.057) | (0.087) | (0.150) | (0.101) |
| Mean Y | 0.154 | 0.160 | 0.160 | -0.731 | 0.321 |
| Observations | 7372 | 5361 | 5361 | 2369 | 2986 |
| *Economic Sufficiency Principal Component, age 40* | | | | | |
| Head Start | -0.113 | -0.093 | -0.082 | -0.219 | -0.127 |
| | (0.090) | (0.106) | (0.131) | (0.180) | (0.155) |
| Other preschool | 0.209** | 0.173 | 0.091 | -0.291 | 0.183 |
| | (0.086) | (0.113) | (0.145) | (0.296) | (0.167) |
| Mean Y | 0.026 | 0.032 | 0.032 | -0.968 | 0.199 |
| Observations | 4085 | 2845 | 2503 | 1065 | 1435 |
| *Good Health Principal Component, Age 30* | | | | | |
| Head Start | -0.248*** | -0.228*** | -0.073 | 0.057 | -0.159 |
| | (0.047) | (0.052) | (0.121) | (0.131) | (0.208) |
| Other preschool | 0.070** | 0.069* | 0.063 | 0.033 | 0.083 |
| | (0.031) | (0.037) | (0.063) | (0.137) | (0.069) |
| Mean Y | 0.003 | 0.013 | 0.013 | -0.309 | 0.062 |
| Observations | 4749 | 3600 | 3114 | 1150 | 1959 |
| *Good Health Principal Component, Age 40* | | | | | |
| Head Start | -0.143 | -0.126 | -0.101 | 0.044 | -0.174 |
| | (0.107) | (0.128) | (0.200) | (0.200) | (0.400) |
| Other preschool | 0.101 | 0.077 | 0.121 | 0.288 | 0.062 |
| | (0.089) | (0.110) | (0.104) | (0.221) | (0.117) |
| Mean Y | 0.009 | 0.015 | 0.015 | -0.259 | 0.056 |
| Observations | 2228 | 1673 | 1306 | 511 | 795 |

Notes:

Table C.3: Effect of Head Start on Inputs to Economic Sufficiency Index at age 30 (PSID)

| | All | Sibs | Mom FE | Blk, FE | Wht, FE |
|---|---|---|---|---|---|
| *High School Graduate* | | | | | |
| Head Start | 0.007 | -0.002 | -0.011 | -0.024 | -0.015 |
| | (0.018) | (0.021) | (0.026) | (0.031) | (0.045) |
| Other preschool | -0.002 | -0.008 | 0.036* | -0.012 | 0.046* |
| | (0.011) | (0.014) | (0.021) | (0.048) | (0.024) |
| Mean Y | 0.913 | 0.912 | 0.912 | 0.862 | 0.921 |
| Observations | 7372 | 5361 | 5361 | 2369 | 2986 |
| *Attended Some College* | | | | | |
| Head Start | 0.038 | 0.039 | 0.046 | -0.016 | 0.120** |
| | (0.024) | (0.029) | (0.033) | (0.036) | (0.053) |
| Other preschool | 0.068*** | 0.069*** | 0.034 | -0.011 | 0.043 |
| | (0.019) | (0.023) | (0.039) | (0.046) | (0.047) |
| Mean Y | 0.531 | 0.532 | 0.532 | 0.396 | 0.556 |
| Observations | 7372 | 5361 | 5361 | 2369 | 2986 |
| *Fraction of last 5 yrs not on Food Stamps/SNAP, age 30* | | | | | |
| Head Start | -0.018 | 0.011 | 0.043 | 0.042 | 0.076 |
| | (0.015) | (0.017) | (0.033) | (0.037) | (0.055) |
| Other preschool | -0.003 | 0.007 | -0.019 | -0.019 | -0.015 |
| | (0.007) | (0.009) | (0.018) | (0.047) | (0.019) |
| Mean Y | 0.936 | 0.929 | 0.929 | 0.831 | 0.949 |
| Observations | 4186 | 2805 | 2175 | 887 | 1285 |
| *Never on AFDC/TANF by age 30* | | | | | |
| Head Start | -0.028* | -0.015 | -0.009 | -0.001 | 0.001 |
| | (0.016) | (0.018) | (0.020) | (0.023) | (0.034) |
| Other preschool | 0.022*** | 0.026*** | 0.004 | -0.010 | 0.005 |
| | (0.008) | (0.009) | (0.011) | (0.025) | (0.012) |
| Mean Y | 0.938 | 0.940 | 0.940 | 0.819 | 0.962 |
| Observations | 7372 | 5361 | 5361 | 2369 | 2986 |
| *Fraction of last 5 yrs with positive earnings, age 30* | | | | | |
| Head Start | 0.041*** | 0.035** | 0.061 | 0.026 | 0.088 |
| | (0.015) | (0.017) | (0.038) | (0.034) | (0.072) |
| Other preschool | 0.013 | 0.008 | 0.015 | -0.047 | 0.027 |
| | (0.011) | (0.013) | (0.019) | (0.048) | (0.020) |
| Mean Y | 0.895 | 0.898 | 0.898 | 0.845 | 0.907 |
| Observations | 4378 | 3295 | 2800 | 1054 | 1740 |
| *Mean Inc. Rel. Pov. in last 5 years, age 30* | | | | | |
| Head Start | -29.579*** | -27.953** | -16.953 | 5.860 | -24.477 |
| | (10.548) | (12.160) | (14.369) | (12.890) | (23.499) |
| Other preschool | 42.704** | 46.790*** | -1.326 | -4.147 | 0.923 |
| | (18.606) | (17.411) | (16.118) | (17.769) | (18.924) |
| Mean Y | 385.831 | 385.933 | 385.933 | 224.651 | 412.236 |
| Observations | 5293 | 4068 | 3694 | 1514 | 2175 |
| *Fraction of last 5 yrs no unemployment, age 30* | | | | | |
| Head Start | -0.007 | -0.001 | 0.005 | -0.013 | 0.056 |
| | (0.015) | (0.016) | (0.030) | (0.031) | (0.049) |
| Other preschool | -0.017 | -0.013 | -0.029 | 0.022 | -0.040 |
| | (0.012) | (0.014) | (0.027) | (0.029) | (0.032) |
| Mean Y | 0.854 | 0.850 | 0.850 | 0.807 | 0.857 |
| Observations | 4259 | 3184 | 2670 | 981 | 1683 |

Notes: This table shows estimates from regressions of the inputs to the Economic Sufficiency Index at age 30 on an indicator for participation in Head Start together with control variables described in the text. Estimates are weighted to be representative of 1995 population; see text for details. Standard errors are clustered at 1968 family id in column 1 and on mother id level otherwise. * p < .10, ** p < .05, *** p < .01. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table C.4: Effect of Head Start on Inputs to Economic Sufficiency Index at age 40 (PSID)

| | All | Sibs | Mom FE | Blk, FE | Wht, FE |
|---|---|---|---|---|---|
| *Fraction of last 5 yrs not on Food Stamps/SNAP, age 40* | | | | | |
| Head Start | 0.001 | 0.009 | 0.045 | 0.054 | 0.051 |
| | (0.019) | (0.020) | (0.033) | (0.044) | (0.049) |
| Other preschool | 0.001 | 0.003 | -0.010 | -0.013 | -0.008 |
| | (0.010) | (0.013) | (0.023) | (0.062) | (0.023) |
| Mean Y | 0.957 | 0.957 | 0.957 | 0.866 | 0.971 |
| Observations | 1972 | 1423 | 1213 | 564 | 647 |
| *Never on AFDC/TANF by age 40* | | | | | |
| Head Start | 0.008 | 0.022 | -0.009 | -0.010 | 0.002 |
| | (0.020) | (0.023) | (0.030) | (0.039) | (0.048) |
| Other preschool | 0.016 | 0.019 | 0.018 | -0.034 | 0.025 |
| | (0.010) | (0.012) | (0.021) | (0.062) | (0.021) |
| Mean Y | 0.932 | 0.933 | 0.933 | 0.778 | 0.959 |
| Observations | 4085 | 2845 | 2503 | 1065 | 1435 |
| *Fraction of last 5 yrs with positive earnings, age 40* | | | | | |
| Head Start | 0.026 | 0.022 | 0.021 | 0.073 | -0.180 |
| | (0.031) | (0.038) | (0.062) | (0.053) | (0.130) |
| Other preschool | -0.004 | -0.012 | -0.026 | -0.135*** | 0.003 |
| | (0.027) | (0.033) | (0.051) | (0.052) | (0.060) |
| Mean Y | 0.850 | 0.849 | 0.849 | 0.856 | 0.847 |
| Observations | 1829 | 1369 | 1078 | 445 | 633 |
| *Mean Inc. Rel. Pov. in last 5 years, age 40* | | | | | |
| Head Start | 1.769 | 3.447 | 32.738 | 27.251 | -11.620 |
| | (21.347) | (26.202) | (30.410) | (24.095) | (56.148) |
| Other preschool | 97.953** | 101.861** | 24.513 | 17.035 | 26.140 |
| | (38.986) | (47.085) | (40.157) | (22.343) | (50.412) |
| Mean Y | 436.769 | 434.280 | 434.280 | 234.965 | 466.741 |
| Observations | 2152 | 1613 | 1272 | 540 | 732 |
| *Fraction of last 5 yrs no unemployment, age 40* | | | | | |
| Head Start | -0.003 | -0.022 | -0.028 | -0.033 | -0.046 |
| | (0.022) | (0.027) | (0.047) | (0.056) | (0.083) |
| Other preschool | -0.011 | -0.011 | -0.026 | -0.053 | -0.016 |
| | (0.017) | (0.021) | (0.037) | (0.060) | (0.044) |
| Mean Y | 0.906 | 0.902 | 0.902 | 0.841 | 0.911 |
| Observations | 1825 | 1365 | 1073 | 440 | 633 |
| *Fraction of last 5 yrs owned home, age 40* | | | | | |
| Head Start | -0.022 | -0.024 | 0.045 | -0.058 | 0.070 |
| | (0.049) | (0.056) | (0.056) | (0.054) | (0.121) |
| Other preschool | -0.041 | -0.053 | -0.057 | -0.079 | -0.058 |
| | (0.037) | (0.044) | (0.058) | (0.079) | (0.074) |
| Mean Y | 0.500 | 0.522 | 0.522 | 0.324 | 0.554 |
| Observations | 2292 | 1625 | 1391 | 642 | 747 |

Notes: This table shows estimates from regressions of the inputs to the Economic Sufficiency Index at age 40 on an indicator for participation in Head Start together with control variables described in the text. Estimates are weighted to be representative of 1995 population; see text for details. Standard errors are clustered at 1968 family id in column 1 and on mother id level otherwise. * p < .10, ** p < .05, *** p < .01. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table C.5: Effect of Head Start on Inputs to Good Health Index at age 30 (PSID)

| | All | Sibs | Mom FE | Blk, FE | Wht, FE |
|---|---|---|---|---|---|
| *Fraction of last 5 yrs smoked less than 1 cigarette/day, age 30* | | | | | |
| Head Start | -0.064* | -0.031 | 0.021 | -0.127* | 0.049 |
| | (0.035) | (0.039) | (0.080) | (0.072) | (0.110) |
| Other preschool | -0.017 | 0.017 | -0.011 | -0.181** | 0.012 |
| | (0.021) | (0.024) | (0.052) | (0.091) | (0.056) |
| Mean Y | 0.745 | 0.755 | 0.755 | 0.785 | 0.750 |
| Observations | 2267 | 1742 | 1174 | 376 | 796 |
| *Fraction of last 5 yrs reported good or better health, age 30* | | | | | |
| Head Start | -0.001 | 0.001 | 0.042 | 0.047 | 0.039 |
| | (0.012) | (0.013) | (0.031) | (0.034) | (0.052) |
| Other preschool | 0.008 | 0.004 | 0.005 | -0.009 | 0.010 |
| | (0.008) | (0.010) | (0.016) | (0.035) | (0.017) |
| Mean Y | 0.948 | 0.950 | 0.950 | 0.890 | 0.959 |
| Observations | 3763 | 2806 | 2292 | 829 | 1459 |
| *Negative Mean BMI in last 5 years, age 30* | | | | | |
| Head Start | -1.063** | -0.982* | -0.485 | 1.408 | -1.514 |
| | (0.436) | (0.506) | (0.765) | (0.984) | (1.128) |
| Other preschool | 0.046 | -0.096 | -0.332 | -0.357 | -0.202 |
| | (0.266) | (0.313) | (0.441) | (1.069) | (0.472) |
| Mean Y | -26.569 | -26.615 | -26.615 | -28.826 | -26.267 |
| Observations | 3248 | 2528 | 1978 | 689 | 1286 |

Notes: This table shows estimates from regressions of the inputs to the Good Health Index at age 30 on an indicator for participation in Head Start together with control variables described in the text. Estimates are weighted to be representative of 1995 population; see text for details. Standard errors are clustered at 1968 family id in column 1 and on mother id level otherwise. * p < .10, ** p < .05, *** p < .01. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table C.6: Effect of Head Start on Inputs to Good Health Index at age 40 (PSID)

| | All | Sibs | Mom FE | Blk, FE | Wht, FE |
|---|---|---|---|---|---|
| *Fraction of last 5 yrs smoked less than 1 cigarette/day, age 40* | | | | | |
| Head Start | -0.022 | 0.013 | 0.002 | 0.074 | 0.099 |
| | (0.047) | (0.050) | (0.075) | (0.077) | (0.148) |
| Other preschool | 0.003 | 0.041 | -0.033 | 0.218** | -0.104 |
| | (0.039) | (0.047) | (0.126) | (0.097) | (0.150) |
| Mean Y | 0.738 | 0.728 | 0.728 | 0.713 | 0.731 |
| Observations | 1280 | 930 | 698 | 300 | 398 |
| *Fraction of last 5 yrs reported good or better health, age 40* | | | | | |
| Head Start | 0.010 | 0.008 | 0.013 | 0.021 | 0.002 |
| | (0.034) | (0.039) | (0.059) | (0.061) | (0.144) |
| Other preschool | 0.016 | 0.010 | 0.026 | 0.026 | 0.025 |
| | (0.029) | (0.035) | (0.023) | (0.065) | (0.023) |
| Mean Y | 0.919 | 0.922 | 0.922 | 0.871 | 0.930 |
| Observations | 1463 | 1116 | 884 | 398 | 486 |
| *Negative Mean BMI in last 5 years, age 40* | | | | | |
| Head Start | -1.218** | -1.297* | -0.976 | -0.475 | 0.501 |
| | (0.613) | (0.731) | (0.867) | (1.055) | (1.251) |
| Other preschool | -0.330 | -0.741 | -1.861*** | 1.271 | -2.360*** |
| | (0.424) | (0.518) | (0.647) | (1.503) | (0.693) |
| Mean Y | -27.504 | -27.433 | -27.433 | -29.491 | -27.095 |
| Observations | 2037 | 1486 | 1116 | 413 | 703 |

Notes: This table shows estimates from regressions of the inputs to the Good Health Index at age 40 on an indicator for participation in Head Start together with control variables described in the text. Estimates are weighted to be representative of 1995 population; see text for details. Standard errors are clustered at 1968 family id in column 1 and on mother id level otherwise. * p < .10, ** p < .05, *** p < .01. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table C.7: Impact of Head Start on High School, College, Earnings, and Criminal Behavior (PSID)

| | All | Sibs | Mom FE | Blk, FE | Wht, FE |
|---|---|---|---|---|---|
| *A. Completed High School* | | | | | |
| Head Start | 0.007 | -0.002 | -0.011 | -0.024 | -0.015 |
| | (0.018) | (0.021) | (0.026) | (0.031) | (0.045) |
| Other preschool | -0.002 | -0.008 | 0.036* | -0.012 | 0.046* |
| | (0.011) | (0.014) | (0.021) | (0.048) | (0.024) |
| R-Squared | 0.098 | 0.105 | 0.028 | 0.050 | 0.038 |
| Observations | 7372 | 5361 | 5361 | 2369 | 2986 |
| *B. Completed Some College* | | | | | |
| Head Start | 0.038 | 0.039 | 0.046 | -0.016 | 0.120** |
| | (0.024) | (0.029) | (0.033) | (0.036) | (0.053) |
| Other preschool | 0.068*** | 0.069*** | 0.034 | -0.011 | 0.043 |
| | (0.019) | (0.023) | (0.039) | (0.046) | (0.047) |
| R-Squared | 0.213 | 0.233 | 0.050 | 0.056 | 0.057 |
| Observations | 7372 | 5361 | 5361 | 2369 | 2986 |
| *C. Ln Earnings 23-25* | | | | | |
| Head Start | 0.040 | 0.032 | 0.064 | 0.057 | 0.113 |
| | (0.056) | (0.066) | (0.109) | (0.142) | (0.158) |
| Other preschool | 0.064 | 0.035 | 0.084 | 0.174 | 0.070 |
| | (0.045) | (0.052) | (0.098) | (0.173) | (0.110) |
| R-Squared | 0.151 | 0.161 | 0.131 | 0.095 | 0.152 |
| Observations | 4351 | 3309 | 2726 | 986 | 1736 |
| *D. Not Booked/Charged with Crime* | | | | | |
| Head Start | -0.007 | -0.012 | -0.008 | 0.028 | -0.068 |
| | (0.025) | (0.031) | (0.033) | (0.028) | (0.064) |
| Other preschool | -0.006 | 0.007 | -0.002 | -0.022 | 0.002 |
| | (0.014) | (0.017) | (0.033) | (0.036) | (0.039) |
| R-Squared | 0.055 | 0.062 | 0.089 | 0.074 | 0.106 |
| Observations | 5005 | 3591 | 3206 | 1366 | 1836 |

Notes: This table shows estimates from regressions of high school graduation (panel A), some college attainment (panel B), ln earnings between ages 23 and 25 (panel C) and not being charged with a crime (panel D) on an indicator for participation in Head Start together with control variables described in the text. Among the 7,372 individuals in the sample, 1098 individuals are in families that have variation in the Head Start variable (347 families), among those for whom we observe completed education; 887 black (277 black families), and 211 white individuals (70 white families). Crime sample limited to individuals age $\geq$ 16 at the time of interview in 1995. Estimates are weighted to be representative of 1995 population; see text for details. Standard errors are clustered at 1968 family id in column 1 and on mother id level otherwise. * p < .10, ** p < .05, *** p < .01. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table C.8: Impact of Head Start on Main Outcomes by Sex

| | Black | | White | |
|---|---|---|---|---|
| | Female | Male | Female | Male |
| *High School* | | | | |
| Head Start x Sex | 0.008 | -0.062 | 0.005 | -0.043 |
| | (0.033) | (0.042) | (0.059) | (0.054) |
| P-value of Difference | 0.092 | | 0.497 | |
| *Some College* | | | | |
| Head Start x Sex | -0.012 | -0.021 | 0.102 | 0.145*** |
| | (0.044) | (0.045) | (0.074) | (0.053) |
| P-value of Difference | 0.873 | | 0.582 | |
| *Ln Earnings 23-25* | | | | |
| Head Start x Sex | 0.265 | -0.238 | 0.133 | 0.078 |
| | (0.171) | (0.202) | (0.217) | (0.174) |
| P-value of Difference | 0.037 | | 0.834 | |
| *No Crime* | | | | |
| Head Start x Sex | 0.038 | 0.016 | -0.036 | -0.112 |
| | (0.035) | (0.041) | (0.073) | (0.089) |
| P-value of Difference | 0.661 | | 0.448 | |
| *Economic Sufficiency Index, age 30* | | | | |
| Head Start x Sex | -0.052 | -0.197** | -0.099 | 0.078 |
| | (0.099) | (0.090) | (0.112) | (0.141) |
| P-value of Difference | 0.148 | | 0.252 | |
| *Economic Sufficiency Index, age 40* | | | | |
| Head Start x Sex | -0.021 | -0.363** | 0.058 | -0.271 |
| | (0.173) | (0.164) | (0.140) | (0.184) |
| P-value of Difference | 0.098 | | 0.099 | |
| *Good Health Index, Age 30* | | | | |
| Head Start x Sex | 0.042 | -0.004 | -0.198 | -0.361 |
| | (0.159) | (0.218) | (0.278) | (0.378) |
| P-value of Difference | 0.838 | | 0.690 | |
| *Good Health Index, Age 40* | | | | |
| Head Start x Sex | 0.349 | -0.672** | 0.605 | -1.099** |
| | (0.273) | (0.271) | (0.378) | (0.480) |
| P-value of Difference | 0.014 | | 0.004 | |

This table shows estimates from regressions of our main outcomes on an indicator for participation in Head Start interacted with an indicator for being female or male. The estimated interactions between Head Start and female (male) are shown in columns 1 and 3 (2 and 4). Estimates are weighted to be representative of 1995 population; see text for details. Standard errors are clustered at 1968 family id in column 1 and on mother id level otherwise. * p < .10, ** p < .05, *** p < .01. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table C.9: Regression: Interaction with Cohort (Linear) (PSID)

| | All | Sibs | Mom FE | Blk, FE | Wht, FE |
|---|---|---|---|---|---|
| *Economic Sufficiency Index, age 30* | | | | | |
| Head Start | -0.054 | -0.033 | -0.038 | -0.081 | 0.094 |
| | (0.066) | (0.073) | (0.086) | (0.104) | (0.153) |
| Head Start x trend | -0.010** | -0.010* | -0.009 | -0.007 | -0.017 |
| | (0.005) | (0.006) | (0.007) | (0.008) | (0.013) |
| Mean Y | 0.094 | 0.096 | 0.096 | -0.552 | 0.213 |
| Observations | 7372 | 5361 | 5361 | 2369 | 2986 |
| *Economic Sufficiency Index, age 40* | | | | | |
| Head Start | -0.042 | -0.038 | -0.030 | -0.155 | -0.026 |
| | (0.084) | (0.093) | (0.104) | (0.136) | (0.118) |
| Head Start x trend | -0.014 | -0.015 | -0.031* | -0.050** | -0.029 |
| | (0.012) | (0.013) | (0.017) | (0.025) | (0.019) |
| Mean Y | 0.020 | 0.025 | 0.025 | -0.670 | 0.142 |
| Observations | 4085 | 2845 | 2503 | 1065 | 1435 |
| *Good Health Index, Age 30* | | | | | |
| Head Start | -0.318*** | -0.291*** | -0.113 | -0.087 | 0.018 |
| | (0.064) | (0.065) | (0.161) | (0.167) | (0.293) |
| Head Start x trend | -0.004 | -0.004 | -0.007 | 0.034** | -0.044 |
| | (0.007) | (0.007) | (0.019) | (0.017) | (0.034) |
| Mean Y | 0.004 | 0.017 | 0.017 | -0.357 | 0.074 |
| Observations | 4749 | 3600 | 3114 | 1150 | 1959 |
| *Good Health Index, Age 40* | | | | | |
| Head Start | -0.135 | -0.110 | -0.129 | 0.066 | 0.422 |
| | (0.149) | (0.167) | (0.210) | (0.188) | (0.513) |
| Head Start x trend | -0.028 | -0.034 | -0.026 | 0.067 | -0.186** |
| | (0.024) | (0.026) | (0.037) | (0.044) | (0.083) |
| Mean Y | 0.011 | 0.015 | 0.015 | -0.290 | 0.062 |
| Observations | 2228 | 1673 | 1306 | 511 | 795 |

Notes: This table shows estimates from regressions of the Economic Sufficiency and Good Health Indices on an indicator participation in Head Start interacted with a normed linear trend in year of birth (year of birth minus 1966, where 1966 represents the first year that Head Start was available). Estimates are weighted to be representative of 1995 population; see text for details. Standard errors are clustered at 1968 family id in column 1 and on mother id level otherwise. * $p < .10$, ** $p < .05$, *** $p < .01$. Source: Panel Study of Income Dynamics, 1968-2011 waves.

# D Appendix: Replication of GTC (2002)

## D.1 Summary

In this appendix we describe the results of our replication of Garces, Thomas and Currie (2002) (GTC). We describe our replication methods is in the Section D.2.

Table D.1 below shows the summary statistics corresponding to Table 1 of GTC for our sample. We include GTC Table 1 for comparison as Table D.2. In general, the results across the two tables are similar, albeit not identical. The most notable difference is that we find a lower share of respondents participate in Head Start, although the difference is smaller for the sibling sample. The shares of respondents who graduate high school and college are higher in our sample than in GTC. We report average earnings from age 23-25 in nominal terms as well as adjusted to 1999 dollars. Our adjusted earnings are consistently higher than GTC's reported adjusted earnings, but our unadjusted earnings are quite close to their mean adjusted earnings. We suspect that GTC may have reported unadjusted earnings, although it is also possible that the discrepancy is due to a slightly larger sample of individuals with earnings in GTC's sample. Again, the number of observations we report in the final row of the table is based on the number of individuals responding to the Head Start participation question.

Our replication of the main regression results in GTC are shown in Table D.3. We include GTC's Table 2 as Table D.4 for comparison. Our regression results are qualitatively similar, especially for the larger samples (panels A, B, and C). GTC found few statistically significant results, one of which was a negative effect of Head Start on high school completion before including controls. We, too, find this negative and significant result, though ours is slightly smaller. The result in Column (6), which GTC find to be positive and significant, we do not find to be significant. Our results for the college outcomes are aligned with the findings in GTC. The magnitudes that we report are not statistically different from GTC and in particular we replicate the key finding that Head Start influences college going for white children and not for black children. Our replication of Panel C is qualitatively similar to GTC. We do not find a statistically significant decrease in black crime rates as GTC do, although our point estimates are consistently negative for blacks. Otherwise, our estimates are quite imprecise and not statistically different from GTC's.

Our earnings results (panel C replication) are quite different from GTC, but this may be due to differences in how we defined earnings rather than differences in our samples. This is apparent in the fact that we have many fewer observations than GTC beginning from column 2 onward, about 24% smaller in column 2 and 48% smaller in column 8.

## D.2 Replication Methodology

This section documents the process of replicating Garces, Thomas and Currie (2002) (GTC) for future scholars wishing to repeat our steps. We describe three stages of the replication: construction

of the dataset, iterations to identify the likely variable definitions, and our final decisions based on these iterations. We also include information about the mechanics of downloading the data and the variables we use.

### D.2.1 Construction of Dataset

We begin by assembling data from the Panel Study of Income Dynamics (PSID), a nationally representative longitudinal dataset that forms the basis for the analysis in GTC. The PSID consists of the survey responses of household heads and their wives, which compose the annual household-level datasets ("family files"), as well as a smaller database of responses of all individuals in the household to a small set of questions ("cross-year individual files"). We merge the family files to the cross-year individual files using the "case id" number, which is present both on the individual and family files. We also merge responses of an individual's mother and father from the crossyear file for those individuals whose mother or father have been identified in the PSID crossyear file.

The result is a dataset with 71,285 individual observations, each of which contains the personal responses of an individual over time, the responses (usually given by the head of household) to the family interview questions for each year, and the responses of an individual's parents to the cross-year survey. The base dataset includes the Survey of Economic Opportunity "poverty oversample" and the Latino oversample, two populations specifically targeted by the PSID in order to improve the representativeness of the survey. We proceed by excluding the Latino oversample in accordance with GTC's footnote 4.

Next, we construct the variables needed to define our sample. GTC delineate the specifications for their sample throughout the paper, and in particular we rely on their descriptions in Section II and footnote 7. A key stratifying variable in GTC is race, which is also a limiting factor for the sample size since the GTC sample is restricted to only black and white individuals (see footnote 4 of GTC). Unfortunately, the PSID does not assign a race to each individual, so race must be imputed from the annual family responses about race. Specifically, the PSID surveys families about the race of the head and wife of the head of household, so an individual's race can only be identified if that individual becomes a head of household or his wife. Otherwise we must infer the race of the individual through their relation to the head of household or his wife.

The process of identifying race from the responses of other family members can be done at any age and from a variety of different family members, so we have experimented with using more and less restrictive definitions. We establish five definitions of race based on the relations through which we allow inference and the survey years over which we make the inference. These definitions are summarized over those two dimensions below in Table D.5.

The second limiting criterion is the age of individuals. GTC include respondents aged 18 and over in 1995, which results in a sample of respondents born between 1965 and 1977. They exclude the 1964 and 1965 cohorts. Since this sample restriction can be defined and replicated in a few different ways with PSID variables, we develop three candidate limitations on age and year of birth

for individuals in our sample. We describe the criteria which define these alternative candidates in Table D.6.

The third criterion is to identify sets of siblings within the remaining sample that comprise the "siblings subsample." Since the identification strategy relies on the inclusion of a mother fixed effect, we define siblings as any two individuals who satisfy the race and age criteria for the sample and have the same unique mother identification number. The mother identification number is a combination of a family identifier and a personal identifying number which is assigned by the PSID. Individuals that do not have a mother identification number are excluded from the sibling subsample.

Next, we flag observations from the SEO poverty oversample with the intention of excluding them as GTC do. We ultimately do not exclude these observations because comparisons of the sample statistics with and without the SEO sample make us speculate that the results in GTC were generated from a sample that included the SEO sample.

We construct sample weights using CPS weights to make the sample representative of the 1995 white and African-American populations. Specifically, we collapsed the 1995 CPS weights to age-race-sex cells (year of birth is not available) and merge the cell weight onto each observation of our sample. Then, we divide the cell weight by the number of individuals in that age-race-sex cell who are in our sample and the resulting individual weight is what we use for our analysis.

### D.2.2 Search for identical dataset construction

As mentioned previously, the sample construction criteria are clearly documented in GTC. For some dimensions, we could think of a few ways to define variables and samples in accordance with their descriptions. Therefore, we conducted tests to determine the procedures that would yield a dataset consistent with GTC, as well as to assess the stability of the results.

Our search iterations hinge on four parameters: inclusion or exclusion of the SEO oversample; the algorithm for identifying an individual's race; the criteria for age; and the order in which we dropped observations and weighted the sample. For this last parameter, we weighted the sample before dropping the Latino oversample as well as after. We do not present the results for the variations on this final parameter because the exercise clearly indicated that dropping the Latino oversample best matched GTC's results regardless of how the first three parameters were defined.

Table D.7 below shows the results of our iteration of the summary statistics results for a select set of variables. Our goal was to match the results to Table 1 in GTC, reproduced on the first row of the table. The number of observations we report is for the variable for Head Start participation, although some variables have fewer observations. For example, over half the observations for the income variable are missing. GTC also report one N for each column, although they also likely had fewer observations for variables like income.

Our sample is weighted based on race, gender, and age variables from the CPS, so we expect that the mean values for the weighted PSID sample should be similar to the CPS means. We

include the CPS means for the three variables as a comparison. The definitions for age and race are as described in the previous section. There are a number of conclusions we draw from this table. First, we speculate that the 25.17 percent black reported in GTC is, in fact, 15.17 percent, which is much closer to the CPS means. Second, inclusion of the SEO oversample adds approximately 1,500 observations to our sample and brings us quite close to the size of the sample and sample means reported in GTC.

As we had hoped, moving from iteration to iteration substantially changes the number of observations, which suggest which decisions produced the sample of GTC. For example, holding SEO and age definitions constant, moving from our conservative definition of race (2) to the liberal definition (4) adds approximately 30 to 50 observations, an approximately 1.5 percent increase in sample size. The specification of age is also important for defining the sample size. For example, the movement from row 1, 1, 2 (N=3,286) to 1, 2, 2 (N=3,548) is an eight percent increase, and the subsequent movement to row 1, 3, 2 (N=4,187) is an 18 percent increase.

Despite the variability in sample size, our sample characteristics are not sensitive to the decisions along each of these dimensions. Additionally, while our results for these select variables are at times statistically different from those of GTC, we remain close to the magnitudes that they report. The race, gender, and age means are very similar across the specifications, likely on account of the weighting. The preschool participation and high school graduation rates are nearly identical throughout, especially when we include the SEO oversample. The exception to this pattern is Head Start participation. The SEO oversample increases the share of respondents who were in Head Start to close to nine percent, which is still lower than the 10.57 percent reported in GTC. We were unable to replicate this high incidence of Head Start participation throughout the iteration process, including in iterations not reported here.

We also performed iterations on the regression models from GTC's Table 2. GTC conduct a similar regression for each of four outcome variables: high school graduation, college graduation, crime, and later earnings. The first of these three are fairly similar: they are defined by one variable in the PSID. In this comparison table we only show results for high school graduation. On the other hand, compiling a consistent variable for earnings is trickier. Here we present results for one of our regressions, but in general we were not able to replicate the findings for this outcome variable.

There are eight different models in GTC. The first three are on the full sample, the sibling sample, and the sibling sample with controls. The next five models use mother fixed effects: first on the full sample, then the full sample split by whether the mother was white or black, and finally for the subset of mothers with less than a high school education, also split by race.

Table D.8 shows a comparison of the results. We show iterations on the same three age restrictions as above, as well as race definitions for definitions 4 and 5 as defined in the previous section. For each regression the corresponding result from GTC is shown on the first row.

Our regression results are qualitatively similar, especially for the larger samples (panel A). GTC found few statistically significant results, one of which was a negative effect of Head Start on high

school completion (result A.1). We, too, replicate this negative and significant result, though ours are smaller. As can be noted in result A.4, our models using later earnings were similar to those in the paper. The result in B.4, which GTC find to be positive and significant, we do not find significant. However, all of our replications of this result fall within the confidence interval they use.

Among our various iterations, the results are stable. Only the result in A.1 has a difference of one standard error between estimates, with the rest of these results never straying more than half a standard error from each other.

### D.3 Final dataset restrictions

Given our iteration exercises, our preferred sample definition includes the SEO poverty oversample, uses age definition 1 and uses race definition 5 as explained in the first section of this appendix. Our choice of age and race definitions is appropriate for three reasons. First, they replicate the GTC adequately. Second, they are a reasonable method for a researcher not attempting to replicate findings. Third, they result in large samples, which is important for additional analyses.

### D.4 More on the data

We downloaded the data files from http://simba.isr.umich.edu/Zips/ZipMain.aspx. Table D.9 shows the variables we downloaded.

Table D.1: Replication of Garces, Thomas, Currie (2002) Summary Statistics

| | All | Head Start | No Head Start | Sibling Sample |
|---|---|---|---|---|
| Head Start | 0.0873 | 1 | 0 | 0.103 |
| | (.282) | (0) | (0) | (.304) |
| Other preschool | 0.266 | 0 | 0.291 | 0.281 |
| | (.442) | (0) | (.454) | (.45) |
| Fraction completed hs | 0.851 | 0.752 | 0.860 | 0.854 |
| | (.356) | (.432) | (.347) | (.353) |
| Fraction attended some college | 0.468 | 0.339 | 0.481 | 0.482 |
| | (.499) | (.474) | (.5) | (.5) |
| Avg. Earnings age 23-25 | 18543.5 | 13361.3 | 18962.7 | 20116.3 |
| | (14929) | (12057) | (15062) | (17141) |
| Avg. Earnings age 23-25 (CPI adjusted) | 20367.9 | 14730.7 | 20823.9 | 21734.8 |
| | (15646) | (12950) | (15758) | (17521) |
| Fraction booked/charged with crime | 0.0998 | 0.124 | 0.0975 | 0.106 |
| | (.3) | (.33) | (.297) | (.308) |
| Fraction African-American | 0.150 | 0.619 | 0.105 | 0.162 |
| | (.357) | (.486) | (.307) | (.369) |
| Fraction female | 0.502 | 0.533 | 0.499 | 0.475 |
| | (.5) | (.499) | (.5) | (.5) |
| Age in 1995 | 23.67 | 23.14 | 23.72 | 23.14 |
| | (3.44) | (3.5) | (3.43) | (3.28) |
| Fraction eldest child in family | 0.345 | 0.335 | 0.346 | 0.364 |
| | (.475) | (.472) | (.476) | (.481) |
| Fraction low birth weight | 0.0608 | 0.110 | 0.0553 | 0.0560 |
| | (.239) | (.314) | (.229) | (.23) |
| Mother's yrs education | 11.36 | 10.00 | 11.49 | 11.17 |
| | (2.58) | (2.44) | (2.56) | (2.54) |
| Fraction whose mother completed hs | 0.772 | 0.585 | 0.790 | 0.770 |
| | (.419) | (.493) | (.407) | (.421) |
| Father's yrs education | 11.46 | 9.806 | 11.60 | 11.37 |
| | (3.01) | (2.78) | (2.98) | (3) |
| Fraction whose father completed hs | 0.725 | 0.475 | 0.747 | 0.717 |
| | (.446) | (.5) | (.435) | (.451) |
| Family income (age 3-6) (CPI adjusted) | 48040.3 | 30253.9 | 49699.4 | 48580.8 |
| | (27470) | (15498) | (27756) | (29193) |
| Had a single mother at age 4 | 0.119 | 0.320 | 0.0998 | 0.108 |
| | (.324) | (.467) | (.3) | (.31) |
| Household size at age 4 | 4.659 | 5.109 | 4.616 | 4.831 |
| | (1.81) | (2.18) | (1.76) | (1.71) |
| Observations | 3399 | 552 | 2847 | 1541 |

Notes: Weighted to be representative of 1995 population; see text for details.

Table D.2: GTC Table 1: Summary Statistics

| | All sample | Head Start | Not in Head Start | Sibling Sample |
|---|---|---|---|---|
| Head Start | 0.1057 | 1 | 0 | 0.1089 |
| | (.0053) | (0) | (0) | (.0073) |
| Other preschool | 0.2834 | 0.1333 | 0.3011 | 0.2771 |
| | (.0077) | (.0151) | (.0085) | (.0105) |
| Pct. completed hs | 0.7660 | 0.6465 | 0.7803 | 0.7721 |
| | (.0074) | (.0216) | (.0079) | (.0101) |
| Pct. attended some college | 0.3714 | 0.2508 | 0.3859 | 0.3880 |
| | (.0085) | (.0196) | (.0093) | (.0117) |
| Average earnings between age 23-25 | - | - | - | - |
| | - | - | - | - |
| Average earnings between age 23-25 - CPI adjusted | 17290 | 12100 | 17810 | 17310 |
| | (690) | (670) | (760) | (1000) |
| Pct. booked/charged with crime | 0.0969 | 0.1104 | 0.0953 | 0.1004 |
| | (.0051) | (.00139) | (.0054) | (.0070) |
| Pct. African-American | 0.2517 | 0.7532 | 0.1924 | 0.2285 |
| | (.0074) | (.00192) | (.0078) | (.0098) |
| Pct. female | 0.5149 | 0.5641 | 0.5091 | 0.5075 |
| | (.0085) | (.0220) | (.0093) | (.0117) |
| Age in 1995 | 23.66 | 23.35 | 23.70 | 23.65 |
| | (.06) | (.15) | (.06) | (.08) |
| Pct. eldest child in family | 0.5311 | 0.5089 | 0.5337 | 0.5057 |
| | (.0056) | (.0141) | (.0061) | (.0076) |
| Pct. low birth weight | 0.0699 | 0.1040 | 0.0659 | 0.0669 |
| | (.0037) | (.0124) | (.0038) | (.0056) |
| Mother's yrs education | 12.14 | 11.33 | 12.24 | 12.30 |
| | (.04) | (.09) | (.04) | (.05) |
| Pct. whose mother completed hs | 0.7037 | 0.5552 | 0.7212 | 0.7815 |
| | (.0078) | (.0221) | (.0083) | (.0097) |
| Father's yrs education | 11.60 | 10.19 | 11.76 | 12.23 |
| | (.06) | (.14) | (.06) | (.07) |
| Pct. whose father completed hs | 0.5612 | 0.2638 | 0.5964 | 0.6330 |
| | (.0085) | (.0196) | (.0091) | (.0113) |
| Family income (age 3-6) - CPI adjusted | 46230 | 26620 | 48540 | 47330 |
| | (460) | (580) | (500) | (670) |
| Had a single mother at age 4 | 0.1642 | 0.4035 | 0.1359 | 0.1306 |
| | (.0061) | (.0216) | (.0061) | (.0079) |
| Household size at age 4 | 4.59 | 4.97 | 4.55 | 4.84 |
| | (.03) | (.09) | (.03) | (.04) |
| Observations | 3255 | 489 | 2766 | 1742 |

Table D.3: Replication of Garces, Thomas, Currie (2002) Regressions

| | All | Sibs | Controls | Mom FE | Blk, FE | Wht, FE |
|---|---|---|---|---|---|---|
| *Panel A. High School* | | | | | | |
| Head Start | -0.064* | -0.017 | 0.009 | 0.031 | -0.017 | 0.093 |
| | (0.034) | (0.043) | (0.040) | (0.057) | (0.063) | (0.092) |
| Other Preschool | 0.082*** | 0.076*** | 0.014 | 0.028 | 0.068 | 0.021 |
| | (0.013) | (0.022) | (0.021) | (0.035) | (0.072) | (0.038) |
| Observations | 3399 | 1541 | 1541 | 1541 | 615 | 923 |
| *Panel B. College* | | | | | | |
| Head Start | -0.027 | -0.021 | 0.033 | 0.100* | -0.039 | 0.232** |
| | (0.035) | (0.053) | (0.045) | (0.059) | (0.059) | (0.094) |
| Other Preschool | 0.200*** | 0.219*** | 0.098*** | 0.047 | -0.062 | 0.059 |
| | (0.025) | (0.034) | (0.033) | (0.044) | (0.101) | (0.049) |
| Observations | 3399 | 1541 | 1541 | 1541 | 615 | 923 |
| *Panel C. Earnings* | | | | | | |
| Head Start | -0.139* | -0.142 | -0.056 | -0.041 | 0.427* | -0.322 |
| | (0.074) | (0.108) | (0.113) | (0.191) | (0.245) | (0.261) |
| Other Preschool | 0.067 | -0.023 | -0.125* | -0.013 | 0.286 | -0.017 |
| | (0.062) | (0.072) | (0.074) | (0.116) | (0.448) | (0.118) |
| Observations | 2118 | 972 | 972 | 779 | 236 | 541 |
| *Panel D. No Crime* | | | | | | |
| Head Start | -0.028 | 0.069 | -0.055 | -0.086 | 0.065 | -0.222* |
| | (0.028) | (0.050) | (0.049) | (0.070) | (0.044) | (0.125) |
| Other Preschool | -0.000 | -0.020 | 0.004 | -0.046 | 0.059 | -0.059 |
| | (0.015) | (0.019) | (0.020) | (0.038) | (0.052) | (0.043) |
| Observations | 3387 | 1537 | 1537 | 1535 | 614 | 918 |

Notes: * p < .10, ** p < .05, *** p < .01. Weighted to be representative of 1995 population; see text for details. SE clustered at 1968 family id in column 1 and at mother id level otherwise.

Table D.4: GTC Table 2: Regressions

| | (1) All | (2) Sibs | (3) Controls | (4) Mom FE | (5) Blk, FE | (6) Wht, FE | (7) Blk, l.e. HS | (8) Wht, l.e. HS |
|---|---|---|---|---|---|---|---|---|
| *Completed high School* | | | | | | | | |
| Head Start | -0.089 | -0.075 | 0.006 | 0.037 | -0.025 | 0.203 | 0.000 | 0.283 |
| | (0.026) | (0.035) | (0.034) | (0.053) | (.065) | (0.098) | (0.071) | (0.119) |
| Other Preschool | 0.085 | 0.073 | 0.003 | -0.032 | -0.056 | -0.014 | -0.080 | -0.019 |
| | (0.016) | (0.022) | (0.021) | (0.038) | (0.064) | (0.048) | (0.077) | (0.067) |
| Difference | -0.174 | -0.148 | 0.003 | 0.069 | 0.031 | 0.217 | 0.081 | 0.302 |
| S.E Difference | 0.028 | 0.037 | 0.039 | 0.062 | 0.085 | 0.105 | 0.097 | 0.126 |
| N | 3255 | 1742 | 1742 | 1742 | 706 | 1036 | 554 | 677 |
| *Attended Some College* | | | | | | | | |
| Head Start | -0.038 | -0.016 | 0.075 | 0.092 | 0.023 | 0.281 | 0.031 | 0.276 |
| | (0.023) | (0.033) | (0.033) | (0.056) | (.066) | (0.108) | (0.067) | (0.120) |
| Other Preschool | 0.142 | 0.149 | 0.023 | 0.050 | -0.007 | 0.095 | 0.022 | 0.0103 |
| | (0.019) | (0.027) | (0.026) | (0.040) | (0.064) | (0.052) | (0.072) | (0.068) |
| Difference | -0.180 | -0.165 | 0.052 | 0.042 | 0.030 | 0.186 | 0.009 | 0.173 |
| S.E Difference | 0.028 | 0.040 | 0.041 | 0.065 | 0.085 | 0.115 | 0.092 | 0.127 |
| N | 3255 | 1742 | 1742 | 1742 | 706 | 1036 | 554 | 677 |
| *ln(earnings 23-25)* | | | | | | | | |
| Head Start | -0.034 | 0.053 | 0.170 | 0.194 | 0.073 | 0.566 | 0.051 | 1.004 |
| | (0.090) | (0.116) | (0.117) | (0.257) | (0.321) | (0.459) | (0.357) | (0.516) |
| Other Preschool | 0.173 | 0.174 | 0.002 | 0.079 | -0.087 | 0.146 | 0.124 | 0.136 |
| | (0.063) | (0.086) | (0.082) | (0.171) | (0.287) | (0.219) | (0.341) | (0.306) |
| Difference | -0.207 | -0.122 | 0.167 | 0.115 | 0.160 | 0.420 | -0.073 | 0.868 |
| S.E Difference | 0.104 | 0.138 | 0.144 | 0.302 | 0.420 | 0.504 | 0.482 | 0.548 |
| N | 1383 | 728 | 728 | 728 | 272 | 456 | 216 | 320 |
| *Booked or charged with crime* | | | | | | | | |
| Head Start | 0.023 | 0.041 | 0.012 | -0.053 | -0.116 | 0.122 | -0.126 | 0.058 |
| | (0.018) | (0.026) | (0.026) | (0.039) | (0.045) | (0.077) | (0.050) | (0.095) |
| Other Preschool | -0.017 | - 0.022 | -0.001 | 0.032 | 0.000 | 0.063 | -0.023 | 0.147 |
| | (0.011) | (0.016) | (0.017) | (0.028) | (0.045) | (0.036) | (0.056) | (0.054) |
| Difference | 0.040 | 0.063 | 0.013 | -0.085 | -0.117 | 0.059 | -0.103 | -0.089 |
| S.E Difference | 0.020 | 0.028 | 0.030 | 0.045 | 0.059 | 0.082 | 0.070 | 0.100 |
| N | 3255 | 1742 | 1742 | 1742 | 706 | 1036 | 554 | 677 |

SE in parentheses.

Table D.5: Alternative Definitions of Race

| Defn. | Survey Years | | Relation to Head (or Wife) | | | | |
|---|---|---|---|---|---|---|---|
| | 1995 | 1985-1996 | Head | Wife | Child | Parent | Sibling |
| 1 | X | | X | X | X | | |
| 2 | X | | X | X | X | X | X |
| 3 | | X | X | X | X | | |
| 4 | | X | X | X | X | X | X |
| 5 | | X | X | X | X | X | X |

Table D.6: Candidate limitations on birth year and age

| Defn. | BirthYears | | | Age in 1995 | | |
|---|---|---|---|---|---|---|
| | 1966-1977 | Not 1965, 1978 | No Restriction | >18 | 17-29 | 17-30 |
| 1 | X | | | X | | |
| 2 | | X | | | X | |
| 3 | | | X | | | X |

Table D.7: Iterations for Summary Statistics Table

| | | | Black | Female | Age | Head Start | Preschool | High School | N |
|---|---|---|---|---|---|---|---|---|---|
| GTC(2002) | | | 0.252 | 0.515 | 23.660 | 0.106 | 0.283 | 0.766 | 3255 |
| CPS 1995 | | | 0.150 | 0.505 | 23.686 | | | | |
| *Sample Iterations* | | | | | | | | | |
| SEO | Age | Race | | | | | | | |
| 0 | 1 | 2 | 0.149 | 0.497 | 22.952 | 0.078 | 0.302 | 0.822 | 1708 |
| 0 | 1 | 4 | 0.149 | 0.497 | 22.950 | 0.079 | 0.299 | 0.820 | 1735 |
| 0 | 2 | 2 | 0.154 | 0.499 | 22.859 | 0.079 | 0.309 | 0.811 | 1855 |
| 0 | 2 | 4 | 0.154 | 0.499 | 22.857 | 0.080 | 0.306 | 0.809 | 1883 |
| 0 | 3 | 2 | 0.150 | 0.503 | 23.713 | 0.076 | 0.286 | 0.820 | 2173 |
| 0 | 3 | 4 | 0.150 | 0.503 | 23.712 | 0.076 | 0.284 | 0.818 | 2204 |
| 1 | 1 | 2 | 0.153 | 0.498 | 22.959 | 0.089 | 0.290 | 0.788 | 3286 |
| 1 | 1 | 4 | 0.153 | 0.498 | 22.958 | 0.089 | 0.288 | 0.787 | 3333 |
| 1 | 2 | 2 | 0.157 | 0.500 | 22.926 | 0.087 | 0.292 | 0.782 | 3548 |
| 1 | 2 | 4 | 0.157 | 0.500 | 22.925 | 0.087 | 0.290 | 0.781 | 3597 |
| 1 | 3 | 2 | 0.150 | 0.503 | 23.710 | 0.082 | 0.276 | 0.788 | 4187 |
| 1 | 3 | 4 | 0.120 | 0.503 | 23.710 | 0.082 | 0.274 | 0.787 | 4244 |

Notes: First row corresponds to selections from Garces, Thomas and Currie (2002) table 1. Second row corresponds to 1995 CPS means, as described in the text of the appendix. The next 12 columns correspond to sample iterations on three criteria. The first is the inclusion (SEO=1) or exclusion (SEO=0) of the Survey of Economic Opportunity sample. The three age criteria and two race criteria are explained in detail in the previous table. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table D.8: Iterations for Regressions Table

| *Panel A.* | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | HS, All | | | HS, Sib | | | HS, Mom FE | | | Log Earnings, All | |
| | | b | se | N | b | se | N | b | se | N | b | se | N |
| GTC (2002) | | -0.089 | (0.026) | 3255 | -0.075 | (0.035) | 1742 | 0.037 | (0.053) | 1742 | -0.034 | (0.090) | 1383 |
| *Sample Iterations* | | | | | | | | | | | | | |
| Age | Race | | | | | | | | | | | | |
| 1 | 4 | -0.075 | (0.030) | 3315 | -0.035 | (0.043) | 1543 | 0.047 | (0.075) | 1543 | -0.064 | (0.106) | 894 |
| 1 | 5 | -0.071 | (0.030) | 3344 | -0.025 | (0.042) | 1565 | 0.047 | (0.075) | 1565 | -0.067 | (0.105) | 898 |
| 2 | 4 | -0.073 | (0.030) | 3585 | -0.034 | (0.039) | 1731 | 0.072 | (0.077) | 1731 | -0.064 | (0.104) | 894 |
| 2 | 5 | -0.067 | (0.031) | 3616 | -0.024 | (0.039) | 1753 | 0.072 | (0.076) | 1753 | -0.067 | (0.104) | 898 |
| 3 | 4 | -0.052 | (0.026) | 4233 | -0.046 | (0.035) | 2125 | 0.037 | (0.063) | 2125 | -0.043 | (0.092) | 1132 |
| 3 | 5 | -0.046 | (0.027) | 4264 | -0.036 | (0.035) | 2147 | 0.036 | (0.062) | 2147 | -0.046 | (0.092) | 1136 |

| *Panel B.* | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | HS, Mom FE, Black | | | HS, Mom FE, White | | | HS, Mom<HS, Black | | | HS, Mom<HS, White | | |
| | | b | se | N | b | se | N | b | se | N | b | se | N |
| GTC (2002) | | -0.025 | (0.065) | 706 | 0.203 | (0.098) | 1036 | 0 | (0.071) | 554 | 0.283 | (0.119) | 677 |
| *Sample Iterations* | | | | | | | | | | | | | |
| Age | Race | | | | | | | | | | | | |
| 1 | 4 | -0.030 | (0.058) | 625 | 0.133 | (0.089) | 898 | -0.026 | (0.058) | 586 | 0.152 | (0.099) | 672 |
| 1 | 5 | -0.030 | (0.058) | 625 | 0.133 | (0.088) | 920 | -0.026 | (0.058) | 586 | 0.152 | (0.098) | 692 |
| 2 | 4 | -0.028 | (0.056) | 702 | 0.181 | (0.094) | 1008 | -0.025 | (0.056) | 649 | 0.203 | (0.105) | 759 |
| 2 | 5 | -0.028 | (0.056) | 702 | 0.181 | (0.092) | 1030 | -0.025 | (0.056) | 649 | 0.202 | (0.104) | 779 |
| 3 | 4 | -0.043 | (0.044) | 858 | 0.120 | (0.081) | 1241 | -0.045 | (0.044) | 797 | 0.136 | (0.092) | 961 |
| 3 | 5 | -0.043 | (0.044) | 858 | 0.114 | (0.079) | 1263 | -0.045 | (0.044) | 797 | 0.130 | (0.088) | 981 |

Notes: First row of each panel corresponds to selections from Garces, Thomas and Currie (2002) table 2. The three age criteria and two race criteria are explained in detail in the previous table. Source: Panel Study of Income Dynamics, 1968-2011 waves.

Table D.9: PSID Variables used in the analysis

| Our Variable | PSID Original Variable | Description (derived variable) | Source |
|---|---|---|---|
| id1968 | ER30001 | Family identifier | Indiv. Cross year |
| pernum | ER30002 | Personal identifier | Indiv. Cross year |
| relation1968-relation2001 | ER30003, ER30022, ER30045, ER30069, ER30093, ER30119, ER30140, ER30162 ER30190, ER30219, ER30248 ER30285, ER30315, ER30345, ER30375, ER30401, ER30431, ER30465, ER30500, ER30537, ER30572, ER30608, ER30644 ER30691, ER30735, ER30808, ER33103, ER33203, ER33303, ER33403, ER33503, ER33603 | Relation to head | Indiv. Cross year |
| caseid1968-caseid2011 | ER30020 ER30043 ER30067 ER30091 ER30117 ER30138 ER30160 ER30188 ER30217 ER30246 ER30283 ER30313 ER30343 ER30373 ER30399 ER30429 ER30463 ER30498 ER30535 ER30570 ER30606 ER30642 ER30689 ER30733 ER30806 ER33101 ER33201 ER33301 ER33401 ER33501 ER33601 ER33701 ER33801 ER33901 ER34001 ER34101 ER33601 | Fam. Intervew Number | Indiv. Cross year |
| edu1968-edu2011 | ER30010 ER30052 ER30076 ER30100 ER30126 ER30147 ER30169 ER30197 ER30226 ER30255 ER30296 ER30326 ER30356 ER30384 ER30413 ER30443 ER30478 ER30513 ER30549 ER30584 ER30620 ER30657 ER30703 ER30748 ER30820 ER33115 ER33215 ER33315 ER33415 ER33516 ER33616 ER33716 ER33817 ER33917 ER34020 ER34119 | Yrs. Education | Indiv. Cross year |
| age1995 | ER33204 | Age in 1995 | Indiv. Cross year |
| birthyr1995 | ER33206 | Birthyear in 1995 | Indiv. Cross year |
| headstart1995 | ER33261 | Head Start Response in 1995 | Indiv. Cross year |
| preschool1995 | ER33264 | Preschool Response in 1995 | Indiv. Cross year |
| preschool1995 | ER33266 | Crime Response in 1995 | Indiv. Cross year |
| sex | ER32000 | Sex | Indiv. Cross year |
| momid1968 | ER32009 | Mother's Family ID | Indiv. Cross year |
| mompernum | ER32010 | Mother's Personal ID | Indiv. Cross year |
| dadid1968 | ER32016 | Father's Family ID | Indiv. Cross year |
| dadpernum | ER32017 | Father's Personal ID | Indiv. Cross year |
| birthweight | ER32014 | Birth weight | Indiv. Cross year |

| Our Variable | PSID Original Variable | Description (derived variable) | Source |
|---|---|---|---|
| crime1995 | ER33266 | Committed/Charged with Crime | Indiv. Cross year |
| parityofmom | ER32013 | Parity of mom (Eldest) | Indiv. Cross year |
| h_edu1968-h_edu2011 | V313 V794 V1485 V2197 V2823 V3241 V3663 V4198 V5074 V5647 V6194 V6787 V7433 V8085 V8709 V9395 V11042 V12400 V13640 V14687 V16161 V17545 V18898 V20198 V21504 V23333 ER4158 ER6998 ER9249 ER12222 ER16516 ER20457 ER24148 ER28047 ER41037 ER46981 ER52405 | Education of Head (Mom, Dad Education) | Family Interviews |
| w_edu1968, w_edu1972-w_edu2011 | V246 V2687 V3216 V3638 V4199 V5075 V5648 V6195 V6788 V7434 V8086 V8710 V9396 V11043 V12401 V13641 V14688 V16162 V17546 V18899 V20199 V21505 V23334 ER4159 ER6999 ER9250 ER12223 ER16517 ER20458 ER24149 ER28048 ER41038 ER46982 ER52406 | Education of Wife of Head (Mom Education) | Family Interviews |
| h_sex1968-h_sex2011 | V119 V1010 V1240 V1943 V2543 V3096 V3509 V3922 V4437 V5351 V5851 V6463 V7068 V7659 V8353 V8962 V10420 V11607 V13012 V14115 V15131 V16632 V18050 V19350 V20652 V22407 ER2008 ER5007 ER7007 ER10010 ER13011 ER17014 ER21018 ER25018 ER36018 ER42018 ER47318 | Sex of Head (Single mom) | Family Interviews |
| f_tanf1994-f_tanf2011 | ER3262 ER6262 ER8379 ER11272 ER14538 ER18697 ER22069 ER26050 ER37068 ER43059 ER48381 | Family Received AFDC/TANF last year | Family Interviews |
| f_fs1994-f_fs2011 | ER3059 ER6058 ER8155 ER11049 ER14255 ER18386 ER21652 ER25654 ER36672 ER42691 ER48007 | Family Received Food Stamps last year | Family Interviews |
| h_cigs1986, h_cigs1999-h_cigs2011 | V13442 ER15544 ER19709 ER23124 ER27099 ER38310 ER44283 ER49621 | Cigarettes Per Day of Head | Family Interviews |
| w_cigs1986, w_cigs1999-w_cigs2011 | V13477 ER15652 ER19817 ER23251 ER27222 ER39407 ER45380 ER50739 | Cigarettes Per Day of Wife of Head | Family Interviews |
| h_wlbs1999-h_wlbs2011 | ER15552 ER19717 ER23132 ER38320 ER44293 ER49631 | Weight of Head (BMI) | Family Interviews |
| w_wlbs1999-w_wlbs2011 | ER15660 ER19825 ER23259 ER27232 ER39417 ER45390 ER50749 | Weight of Wife of Head (BMI) | Family Interviews |

| Our Variable | PSID Original Variable | Description (derived variable) | Source |
|---|---|---|---|
| h_srhealth1984-h_srhealth2011 | V10877 V11991 V13417 V14513 V15993 V17390 V18721 V20021 V21321 V23180 ER3853 ER6723 ER8969 ER11723 ER15447 ER19612 ER23009 ER26990 ER38202 ER44175 ER49494 | Self-Reported Health of Head | Family Interviews |
| w_srhealth1984-w_srhealth2011 | V10884 V12344 V13452 V14524 V15999 V17396 V18727 V20027 V21328 V23187 ER3858 ER6728 ER8974 ER11727 ER15555 ER19720 ER23136 ER27113 ER39299 ER45272 ER50612 | Self Reported Health of Head of Wife | Family Interviews |
| f_rentown1968-f_rentown2011 | V103 V593 V1264 V1967 V2566 V3108 V3522 V3939 V4450 V5364 V5864 V6479 V7084 V7675 V8364 V8974 V10437 V11618 V13023 V14126 V15140 V16641 V18072 V19372 V20672 V22427 ER2032 ER5031 ER7031 ER10035 ER13040 ER17043 ER21042 ER25028 ER36028 ER42029 ER47329 | Family Rents/Owns Home | Family Interviews |
| h_wages1968-h_wages2011 | V251 V699 V1191 V1892 V2493 V3046 V3458 V3858 V4373 V5283 V5782 V6391 V6981 V7573 V8265 V8873 V10256 V11397 V12796 V13898 V14913 V16413 V17829 V20178 V21484 V23323 ER4140 ER6980 ER9231 ER12080 ER16463 ER20443 ER24116 ER27931 ER40921 ER46829 ER52237 | Earnings of Head | Family Interviews |
| w_wages1968-w_wages2011 | V76 V516 V1198 V1899 V2500 V3053 V3465 V3865 V4379 V5289 V5788 V6398 V6988 V7580 V8273 V8881 V10263 V11404 V12803 V13905 V14920 V16420 V17836 V19136 V20436 V23324 ER4144 ER6984 ER9235 ER12082 ER16465 ER20447 ER24135 ER27943 ER40933 ER46841 ER52249 | Earnings of Wife of Head | Family Interviews |

# E Functional form choices with Binary Treatment and Binary Outcome

We now consider potential sensitivity to functional form modeling assumptions. For binary outcomes the usual choice of specifications include linear probability model (LPM), logit, and probit. In the cross-sectional setting, the conventional wisdom is that the choice among these options is fairly innocuous, especially when the objective is to recover the ATE.[51] We are not aware of any previous systematic exploration of these properties in extremely short-panel settings such as found in the FFE design. We demonstrate some complications that arise in such settings, and compare the performance of these estimators.

## E.1 Specification choices

Empiricists commonly use LPM specification to estimate FE models. In our sample of papers, this is almost universally used as the primary, if not only, specification. We speculate that this is motivated by (1) the intuition carried over from the cross-sectional case that LPM models usually recover the ATE; (2) the benefit that the incidental parameters problem does not pollute the main parameters of interest (Chamberlain, 1980);[52] (3) computational ease, especially when paired with other complications to the research design such as many fixed effects, instrumental variables, etc.); and (4) the fact that the estimated coefficient $\beta_{LPM}$ directly gives the estimate of the ATE.

Obtaining ATE from a nonlinear specification is not only less common, but also sometimes less straightforward. The conditional logit model, sometimes referred to as logit FE, consistently estimates $\beta_{Logit}$ by conditioning on the number of successes in a family, but does not have a paired method for obtaining treatment effects. To obtain ATE, Wooldridge (2010, section 15.8) recommends employing a regular logit model and including family-level-means of control variables, i.e. "Chamberlain-Mundlak controls," (hereafter, Mundlak controls) rather than directly controlling for fixed effects (Mundlak, 1978; Chamberlain, 1980).[53]

Fernandez-Val (2009) examines the probit FE model. He proposes a bias-correction approach, which is based on the large-T asymptotic bias resulting from the incidental parameters problem. He also derives a "small bias" property for uncorrected/naive estimates of marginal effects for the probit FE model, and demonstrates this for panels of length as short as T=4. However it is not clear that the results in Fernandez-Val (2009) should apply in the family FE setting. This is because: (1)

[51] See Angrist and Pischke (2009, pg. 107) and Wooldridge (2010, section 15.6). In contrast, Cameron and Trivedi (2005, pg. 471) recommend limiting LPM's to exploratory analysis, and note that it does not do a good job making predicted probabilities for individual observations. In panel contexts, textbook treatments generally state that estimates should be fine using LPM (Wooldridge, 2010, pg. 608).

[52] Because this inconsistency is based on the panel length being fixed, the problem may be especially acute for short panels.

[53] The traditional implementation is to model the residual variance as having an i-level random effect, hence the terminology Correlated Random Effects given to this method. However, it is also possible to include family means of control variables and then estimate regular pooled logit or probit, as we will do.

we face extremely short panel lengths due to families commonly having only 2 children[54]; (2) his results apply only to the leading (order 1/T) bias term, but with very short panels the subsequent bias terms could still be relevant; and (3) there is an unresolved challenge of how to address the extrapolation from the estimation sample to singletons (when they are in the target population).[55]

Mundlak controls and naive-fixed-effects methods have the attractive properties of: (1) being easy to implement; (2) respecting the binary functional form of the left-hand-side (LHS) variable; and (3) straightforwardly obtaining ATEs. Nonetheless, empiricists' use of either of these options is uncommon; in our sample of 58 papers discussed in section 2 these methods are not used.

An additional complication with conditional or fixed effect logit and probit models is that they use less variation relative to LPM. With these models, for any families that have no variation in outcomes, i.e "all successes" or "all failures", the fixed effect parameters will be driven to +/- infinity, and these families will be dropped from estimation. This leaves only "double switchers": families with variation in both the outcome variable and the treatment variable. This means that moving from LPM to nonlinear specification is automatically tied to a change in estimation sample, which can reduce the effective sample size and may exacerbate the issues discussed in Section 3. In our application for example we see a reduction from 2986 individuals in the overall white "siblings sample" to 211 individuals in the "RHS switchers" sample to 98 individuals (from only 27 families) in the double switchers sample. A related issue is that the LPM results will depend on the fraction of observations in families that are not LHS switchers, whereas the logit model estimates will be invariant to the number of these non-switchers.

### E.2 Obtaining Marginal Effects from Conditional Logit

In order to address the challenge of translating the conditional logit coefficient, $\beta_{Logit}$, into ATE units that can be compared with LPM results, we introduce a "two-step logit" model. The first step is the usual conditional logit estimator, used to obtain a consistent coefficient $\hat{\beta}$ for variables that change within-family. The second step estimates a random effects logit model (over the full sample, including non-switchers), while imposing the coefficient on the treatment variable (and on other individual-level variables) from the first step model. The purposes of the second step are (1) to estimate coefficients on family-level variables, so as (2) to assign an estimated "logit index" value to each observation, and (3) to estimate the variance of the family-level random effect $\sigma_u^2$. After the second step model is estimated, we then estimate the ATE using:

$$ATE_{2StepLogit} = \frac{1}{N}\sum_{i=1}^{N}\int_{u}(\hat{\beta}_{HeadStart}\cdot\Lambda(\hat{\beta}X_{if}+\hat{\gamma}Z_f+u)\cdot 1-\Lambda(\hat{\beta}X_{if}+\hat{\gamma}Z_f+u)\phi(u)du \quad (25)$$

[54] We have reproduced the results for mean bias from his Table 4 for Probit and LPM-FS. We then reduced the panel size to T=2, and we find a detectible bias of -6.4% of the true ATE for the Probit, and no bias for the LPM-FS

[55] For singletons there is no ability to separately identify the value of the fixed effect from the idosyncratic error term. This is not a problem when the target population is either RHS switcher or all siblings. For these target populations, the naive logit FE or probit FE model could be used following the reweighting ideas presented above.

With $\hat{\beta}_{HeadStart}$ the coefficient on Head Start from the conditional logit first step; $\hat{\beta}$ the coefficient on i-level variables $X_{if}$ from the conditional logit first step; $\hat{\gamma}$ the coefficients on family-level variables from the second step; and $\phi(u)$ the PDF from a normal distribution, with variance $\sigma_u^2$ estimated from the second step family-level random effects model. We have not yet found a prior implementation of this estimator in the literature; but it is similar in spirit to the two step fixed-effects logit proposed by Beck (2015).[56]

### E.3 Selection in Nonlinear Models

A desirable feature of the two-step logit and the Mundlak controls models is that both allow the marginal effect of treatment, $\frac{\partial Pr(Y=1)}{\partial HeadStart}$, to vary across individuals. In both cases, the treatment effect depends on the "index value" for each individual. However, the models maintain an assumption of constant treatment effects in logit units ($\beta_{Logit}$). If the model is misspecified, and instead there are variable treatment effects for different individuals, and that a reweighted estimation sample might produce more reliable results, especially when trying to measure the ATE for a pre-specified target group. This consideration is analogous to the treatment effect heterogeneity discussed in section 4.2.

We propose employing the in-regression weights $\widetilde{s_f^{sw \to tg}} \cdot v_f$ as discussed above in section 4.2. That is, the weights are a combination of (1) propensity score weights derived from a multinomial logit model predicting "RHS switcher" status and "in target population" status, and (2) inverse within-family conditional variance of the treatement variable of interest. For expediency, we continue to estimate this conditional variance from a linear model, and to apply it directly to the second stage logit estimation step.

We explore some of these models in the context of our empirical example, and find some differences in the point estimates and precision across linear and nonlinear specifications. Compared with LPM, we find somewhat smaller and less precise impacts of Head Start on some college when we use the 2-step approach (point estimate: 0.086 (se: 0.059). We note that the slight decrease in precision here accompanies many fewer observations, which has fallen to 1200 for estimation of the logit beta instead of 2987 in the LPM.[57] The point estimate for the Mundlak controls is very similar to LPM, 0.126, but the standard errors are 20% larger (se: 0.053), so that the estimate is significant only at the 10% level.

### E.4 Monte Carlo for Nonlinear Specifications

We next consider the bias of the different specifications in the context of a specific data generating process (DGP).

[56] Beck's second step is a logit FE (with dummies) estimator, with the $\beta$ imposed from the conditional logit first stage. Then the estimated fixed effects are used to obtain the ATE.

[57] Note that in the second step, the ATE is calculated over the full population. Another difference is that we weight the conditional logit regressions using family averages of individual weights, since conditional logit does not accomodate individual weights.

For our simulations, we continue with the PSID data setup presented above in Section 4.4. We take the original data, and estimate a logit model predicting some college attainment, using as regressors family level variables and family-level averages of individual variables. From this model we construct a family level logit index variable, $x_f$. For each simulation, the underlying logit index for each individual is equal to $x_f$, plus the Head Start dummy multiplied by the Head Start (logit) treatment effect. We then turn this index into $Pr(y = 1)$ using the logistic CDF, and then randomly draw outcomes $y$. We consider three DGPs. The first of these has a constant treatment effect (in logit units). The second has a treatment effect that is zero for small families (with 2 or 3 children), and a larger treatment effect for families with 4 or more children. The third DGP has a variable treatment effect which is decreasing linearly in $x_f$. For all of the DGPs, the treatment effect in terms of $Pr\left(y = 1\right)$ will vary across target populations because different children have different logit indices. For DGPs 2 and 3, there is additional variability stemming from family characteristics.

We run 2,500 Monte Carlo replications. In each replication we estimate a basic LPM, and LPM reweighted for the target population. We also estimate our two-step logit model and a logit model with Mundlak controls. For each of these we estimate both an unweighted version and a version that is reweighted for the target population. We consider the same four target populations, and present the results in Table E.1. The first panel shows results for DGP 1, with constant (in logit units) treatment effects. For this DGP, all models perform well for target groups of switchers, with biases that are small and usually not distinguishable from zero. When we target siblings, all children, or Head Start participants, the LPM model exhibits a detectable bias, which is slightly reduced by reweighting. The proposed 2-step logit model and Mundlak model do better, with small bias. However when they are reweighted with an aim to be representative of the target population, they too have a detectable bias.

In DGP 2 we now have treatment effects that vary with family size. Here all of the basic models perform poorly, both LPM and our two logit variations. Reweighting helps dramatically here, for all three models.

For DGP 3 all models give biased results when we target all children or all siblings. The three reweighted models perform roughly equally well. Each of the specifications does well for estimating treatment effects for switchers, Head Start participants, and Head Start siblings, with small biases.

In results not reported, we also explored a naive logit fixed effects specification for target groups of RHS switchers and sbilings. For these groups, this method performs similarly to the LPM, 2-step Logit, and Mundlak logit discussed above.

### E.5 Discussion of Specification Choices

In our literature sample, use of OLS/LPM methods is ubiquitous. Based on the results of this section, we recommend continued use of this method. For researchers who want to pursue a logit type specification, we believe that either the two-step logit model (based off of a conditional logit estimation first step) or a logit with Mundlak controls can perform well.

Table E.1: Monte Carlo Experiments: Bias of Linear and Nonlinear Models Relative to Target ATE, and Effectiveness of Reweighting

| | | LPM | | Logit | | Logit Reweight | |
|---|---|---|---|---|---|---|---|
| | True ATE | FE Baseline | Reweight | 2-Step | Mundlak | 2-Step | Mundlak |
| *1. Constant TE* | | | | | | | |
| Switchers | 86.4 | -0.5 | -0.4 | -0.3 | -1.6* | -1.5 | -1.8* |
| Siblings | 78.8 | 7.0* | 5.6* | 2.0* | 0.6 | 5.5* | 4.3* |
| All | 78.8 | 7.1* | 5.8* | 2.0* | 0.7 | 5.7* | 4.3* |
| HS Participants | 88.1 | -2.3* | -2.1* | 1.0 | -0.2 | -2.8* | -1.0 |
| *2. Large Family TE* | | | | | | | |
| Switchers | 79.6 | -10.2* | 0.0 | -11.5* | -10.8* | -2.5* | -0.9 |
| Siblings | 44.5 | 24.9* | 2.6* | -9.1* | 20.0* | 1.1 | 2.0* |
| All | 36.1 | 33.2* | 1.6 | 0.5 | 28.3* | 0.5 | 1.1 |
| HS Participants | 40.1 | 29.2* | -0.6 | 40.7* | 30.9* | -1.7* | -0.1 |
| *3. TE linear in $X_f$* | | | | | | | |
| Switchers | 102.2 | 0.1 | 0.8 | -1.5 | -1.3 | -2.3* | -1.1 |
| Siblings | 84.3 | 18.1* | 9.3* | 3.8* | 10.4* | 7.3* | 7.9* |
| All | 84.2 | 18.2* | 9.5* | 9.6* | 10.5* | 7.6* | 8.0* |
| HS Participants | 101.9 | 0.4 | -0.2 | 2.6* | 2.5* | -2.8* | 0.8 |

Notes: This table shows the results from 2,500 Monte Carlo simulations for three different DGPs of some college attainment, presented separately in each panel of the table, and four different target populations, shown in each row of the panel. The true DGP is a logit model, and is discussed in Section E.4. The first panel shows results where Head Start has a constant treatment effect (TE) (on the logit index) for all individuals; the second shows results where Head Start (HS) has no effect on individuals from small families (3 or fewer children) and a large effect for families with many children (4 or more children); and the third panel shows results where effects are linear in $X_f$. Column 1, "True Beta," presents the true average increase in the probability of completing some college for participants in Head Start in the sample, which is a function of the DGP and sample composition. The remaining columns present the bias of various estimation strategies, defined as the difference between the estimated effects of Head Start and the true beta. Columns 2 and 3, LPM and LPM reweight, are defined as in Table 4. Columns 4 to 7 show the results from using the two step random effects estimator and Mundlak logit without and with propensity score weights, respectively. Reweighted estimates obtained using in-regression weighting, which accounts for the representativeness of switchers and the conditional variance of Head Start within families. All betas are multiplied by 1,000. * $p < .01$.