NBER WORKING PAPER SERIES

SELECTION INTO IDENTIFICATION IN FIXED EFFECTS MODELS, WITH APPLICATION TO HEAD START

Douglas L. Miller
Na'ama Shenhav
Michel Z. Grosz

Working Paper 26174
http://www.nber.org/papers/w26174

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2019

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, Hebrew University, 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. The views expressed in this article are not necessarily those of the Federal Trade Commission, its commissioners, or the National Bureau of Economic Research.

Selection into Identification in Fixed Effects Models, with Application to Head Start
Douglas L. Miller, Na'ama Shenhav, and Michel Z. Grosz
NBER Working Paper No. 26174
August 2019
JEL No. C33,I28,I38,J13

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 2.6 percentage point (p.p.) increase (s.e. = 6.2 p.p.) in the likelihood of attending some college for white Head Start participants in the PSID. This ATE is 78% smaller than the traditional FFE estimate (12 p.p). Reweighting the CNLSY FE estimates to obtain the ATE produces similar attenuation in the estimated impacts of Head Start.

Douglas L. Miller
Department of Policy Analysis
and Management
MVR Hall
Cornell University
Ithaca, NY 14853
and NBER
dlm336@cornell.edu

Na'ama Shenhav
Department of Economics
Dartmouth College
6106 Rockefeller Hall
Hanover, NH 03755
naama.shenhav@dartmouth.edu

Michel Z. Grosz
Bureau of Economics Federal
Trade Commission 600
Pennsylvania Avenue NW
Washington, DC 20580
mgrosz@ftc.gov

Supplementary results and tables are available at http://www.nber.org/data-appendix/w26174

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 5.2.

[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 out of 35 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 2.6 percentage point (p.p.) increase in the likelihood of attending some college for Head Start participants (s.e. = 6.2 p.p.), and a 6.8 p.p. (se = 6.0 p.p.) increase for the Head Start eligible population. The ATE for Head Start participants estimate is 78% 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 91% smaller than the estimated effects on college-going in GTC for this population; 45% to 91% smaller than unadjusted estimates for all participants from other FFE studies (Bauer and Schanzenbach, 2016; Deming, 2009); and 51% 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 44% smaller and not statistically significant. 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.

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.

Further, the prevalence of non-switching 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 and other two-way FE 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); Chaisemartin and D'Haultfoeille (2019). We discuss SI in two-way FE designs in Section 5.2.

[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 treatment may be correlated with group characteristics, e.g. mean family income, but 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 the probability of receiving 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 participation decision within families. If the probability of attending Head Start is a constant, π, and independent across siblings in a family, then the probability of switching, $P(S_g = 1)$ is simply a function of π 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 π, peaking at $\pi = 0.5$. Further, for a given level of π, 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 π, 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 π.

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, $\frac{Pr\left(S_{g(i)}=1|\mathbf{X_{ig}}\right)}{Pr\left(HeadStart_i=1|\mathbf{X_{ig}}\right)}$, which gives a measure of how aligned the characteristics (vector $\mathbf{X_{ig}}$) 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 δ_{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] $Var(D_i | g(i) = g)$ is calculated using the population formula for variance, $Var(D_i | g(i) = g) = \frac{1}{n_g} \sum_{i \in g} \left(D_i - \frac{\sum_{i \in g} \mathbf{1}(D_i = 1)}{n_g} \right)^2$, 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 δ's change and also the ω's change. The change in the δ'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, ω_z, and how much is driven by the change in identification, δ_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 ω_z between OLS and FE, multiplied by an α-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 δ_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 ω_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 (Overlap in P_x):

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

Fourth, we require a positive probability of being a switcher for each value of $\mathbf{X_g}$ in the target group, which ensures that we can use the switcher sample to recover the distribution of treatment effects in the target sample. Since some covariate values may not be observed in the switcher sample, this assumption implicitly places some restrictions on the relationship between treatment effects and these covariates. For example, since we do not observe 1-unit groups ("singletons") in the switcher sample – they are "never-switchers" – but they are present in some of our application target populations,[24] we cannot allow treatment effects for singletons to be outside the support of

[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.

[24] Singletons comprise 6% of Head Start participants in the CNLSY, and 18% of Head Start participants in the PSID.

the treatment effects of switchers. This also precludes us from including an indicator for singletons in $\mathbf{X_g}$.

In our application we preserve overlap by assuming that treatment effects are the same for all groups with 2 or fewer units, which allows us to extrapolate treatment effects for singletons from 2-unit groups.[25] An alternative approach is to extrapolate treatment effects for never-switchers using functional form assumptions, which we discuss under extensions below.

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 $w^t_{g(i)}$,

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

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. The treatment estimate for each switcher group g is weighted proportionately to match the share of the target population with observable characteristics matching g, which gives the ATE under the assumptions above.[26]

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

In order to implement this reweighting strategy, we first need to obtain estimates of P_x and Q_x. When the identifying sample is a subset of the target population, $Q_x = 1$, and P_x can be estimated by a logit or probit model. Otherwise, these elements can be calculated by using a multinomial logit

[25] We implement this by including an indicator for "1 or 2 child families" in $\mathbf{X_g}$ together with indicators for other family sizes. Alternatively, the target group can be defined to only include families that are ever switchers, such as "siblings" or "multi-child Head Start families."

[26] See Appendix A for a simple derivation of the weights.

model to estimate the probability of each of the four possible combinations of having $S_g = 1/S_g = 0$ and $T_g = 1/T_g = 0$. 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. [27]

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 $\widehat{w^t_{g(i)}}$ to the group-level and perform a normalization to obtain the final estimation weights, $\hat{s^t_g} = \frac{\widehat{w^t_{g(i)}} \cdot n_g}{\sum_{g' \in G_s} \widehat{w^t_{g(i)'}} \cdot n_{g'}} = \frac{\frac{Q_x}{P_x} \cdot n_g}{\sum_{g' \in G_S} \frac{Q_x}{P_x} \cdot n_{\prime g}}$. The 2-step ATE combines these using:

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

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

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

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 ω_{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).[28] Then, the ATE can be estimated by $\widehat{\delta^t_{1step}}$ from a one-step regression using $\widehat{w^t_{g(i)}} \cdot v_g$ as regression weights, and computation of cluster-robust standard errors is straightforward.[29]

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

[27] $\widehat{w^t_{g(i)}}$ can also be multiplied by survey weights, as we do in our PSID example.

[28] 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.

[29] 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 δ_g are independent of one another. This enables Equation 15.

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 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.[30] 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 $Y_{ig}^* = X_{ig} + \beta_{ig} HeadStart_{ig}$, where β_{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 sample 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 3,000 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.[31] The FE estimate is the same for all target populations. We consider four target populations: (i) individuals in Head Start switching families;[32] (ii) all siblings; (iii) all individuals in the sample (including

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

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

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

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 3 to 20% 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, 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.[33]

Panels C and D of Table 5 examine 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 (-0.2 p.p, and -0.08 p.p. on a base of 9 p.p.), but has much 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.[34]

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, and when the model for the probability of switching is correctly specified.

5 Extensions

5.1 Projecting Treatment Effects for "Never-Switchers"

As noted above, the reweighting estimator in Proposition 1 only recovers the ATE for the target population if (i) the target does not include never-switchers or (ii) if the treatment effects for never-

[33] In results not reported, we have examined adding X_{ig} 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.

[34] This is because $Pr(S_g = 1)$ is misspecified as a linear function of X_{ig}, which causes us to misassign the weight for each treatment effect.

switchers in the target can be assumed to be the same as some other target groups with $P_x > 0$. Otherwise, the reweighting estimator only obtains the ATE for the subset of the target with $P_x > 0$, for whom treatment effects are identified.

A slight variant of (ii) above which could also enables recovery of the full target ATE is to extrapolate treatment effects for never-switchers. This requires a stronger form of CFEI: that treatment effects are not only a function of observable characteristics, but that the researcher can correctly specify the functional form of this relationship.[35] This assumption may not be warranted if heterogeneity is primarily driven by unobservable characteristics; in cases where there is support in the data for such a relationship with observed covariates (e.g. increasing effects with group size), this may be a reasonable way to proceed. The weighted average of estimated treatment effects for $P_x > 0$ and extrapolated effects for $P_x = 0$ gives the ATE for the target group.

5.2 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.[36] 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 of the effective number of observations.[37] 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,

[35] See Appendix A.1.1 for a formalization of this assumption and an extension of Proposition 1 using exrapolation.

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

[37] 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

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.

5.3 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.

5.4 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

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).

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.[38]

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.[39]

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, 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.[40] 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.[41]

[38] 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.

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

[40] 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.

[41] Appendix Table B.4 includes summary statistics for the inputs to the summary indices. Appendix Tables B.5,

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.[42] δ_g is a mother fixed effect which enables comparisons across siblings with a shared mother. The vector $\mathbf{X_{ig}}$ 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.[43]

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.[44] 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 β_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;[45] (2) all Head

B.6, and B.7 contain the number of observations for each outcome and control variable in the analysis .

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

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

[44] 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.

[45] An individual is considered Head-Start-eligible if at any point between the ages of 2 and 5 her family income was

Start participants; and (3) all siblings.[46] 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 indicators for family size (grouping together 1 and 2 child families).[47] 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 Ignorability

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.[48]

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, employing traditional inverse propensity score weights to balance observables between target and non-target switchers.[49] This test passes with no sign of systematic differences across target and non-target individuals across all outcomes.[50]

Our reweighting procedure also relies on adequate overlap of P_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 4 individuals, 5% of the Head Start non-switcher observations, and 2% of all Head Start participants. We interpret this magnitude

below 150% of the poverty level, to account for our imperfect ability to observe reportable income.

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

[47] Results are similar when we substitute family size indicators with linear and quadratic terms in family size.

[48] 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.

[49] For target individuals the weights are $1/Pr\left[T_i = 1, S_g = 1|X_{ig}\right]$, and for non-target individuals the weights are $1/\Pr[T_i = 0, S_g = 1|X_{ig}]$.

[50] 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.

of violation of the overlap assumption as mild enough to disregard in our subsequent analysis.[51]

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).[52] 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 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 0.068, 0.026, and 0.079, 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 34% to 78% smaller impact on college attendance. Putting these estimates in broader perspective, they are 45 to 91% smaller than the unadjusted estimates for *all* participants from other FFE studies (Bauer and Schanzenbach, 2016; Deming, 2009) and 51% 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

[51] 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.

[52] 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.

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 0.22 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 an 8.5 p.p. increase in high school graduation ($p < 0.01$), a 7.2 p.p. decline in idleness ($p < 0.10$), a 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 44% smaller, and not statistically significant. We also see substantial 24% and 28% declines in the estimated impact on idleness and having a learning disability, respectively, when we consider the impact on participants. The poor health estimates are relatively more stable; the reweighted impacts on participants are just 3% 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.

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 -0.439. In contrast, the coefficients are relatively stable for blacks.[53]

We explore other reweighting strategies in Appendix Tables B.12 and B.13. Reweighting using

[53]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 -0.211 (s.e. = 0.073).

linear extrapolation of treatment effects to singletons in Table B.12 produces qualitatively similar results to the baseline reweighting, although the point estimates for Head Start participants are often smaller.[54] Table B.13 presents the results when we reweight the FFE estimates using sample shares instead of propensity score weights. Across all outcomes, these estimates are quite similar to the FFE estimates, 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.14 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.

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 group size. In the Head Start setting, heterogeneity along these lines creates an upward-bias in the FE

[54] We have also explored excluding singletons altogether from the target. The estimates for non-singleton Head Start participants and non-singleton Head-Start-eligible children typically lie between the reweighted estimates for siblings and Head Start participants.

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 $Pr(DV_g = 1)$ 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 2.6 p.p. increase in the likelihood of attending some college for Head Start participants, and a 6.8 p.p. increase for Head Start eligible. The ATE estimate for participants is 78% 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 (4.8 p.p.) is 44% smaller than the FFE estimate, a difference which is statistically different at the 10% level. We find relatively 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*.

Chaisemartin, Clement, and Xavier D'Haultfoeille. 2019. "Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects." *mimeo*.

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.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

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.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1995. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data." *Journal of the American Statistical Association*, 90(429): 106–121.

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*.

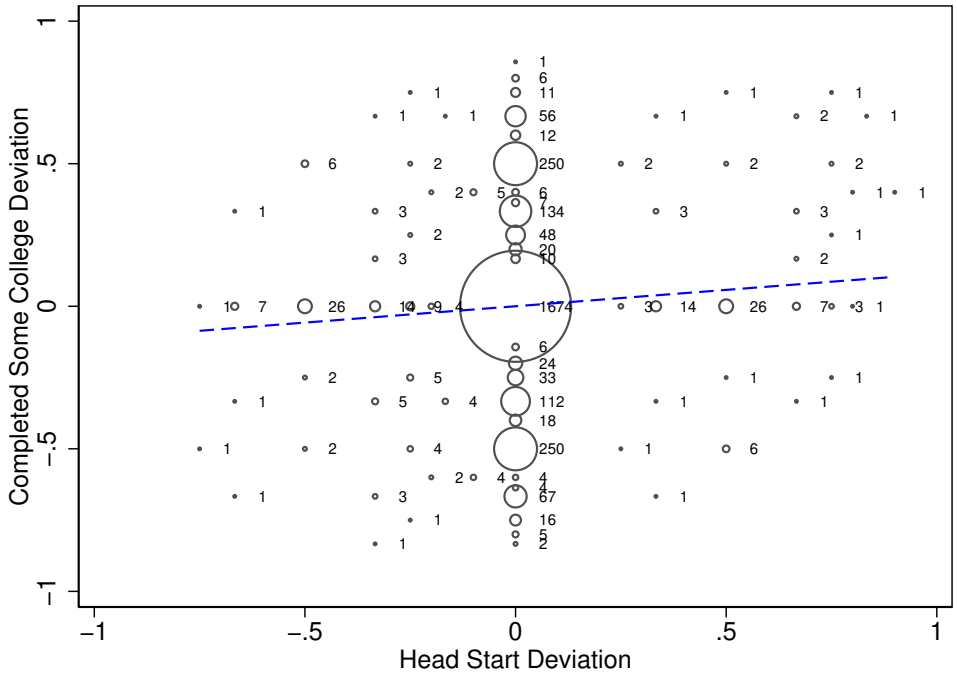
Wiswall, Matthew. 2013. "The dynamics of teacher quality." *Journal of Public Economics*, 100: 61–78.

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

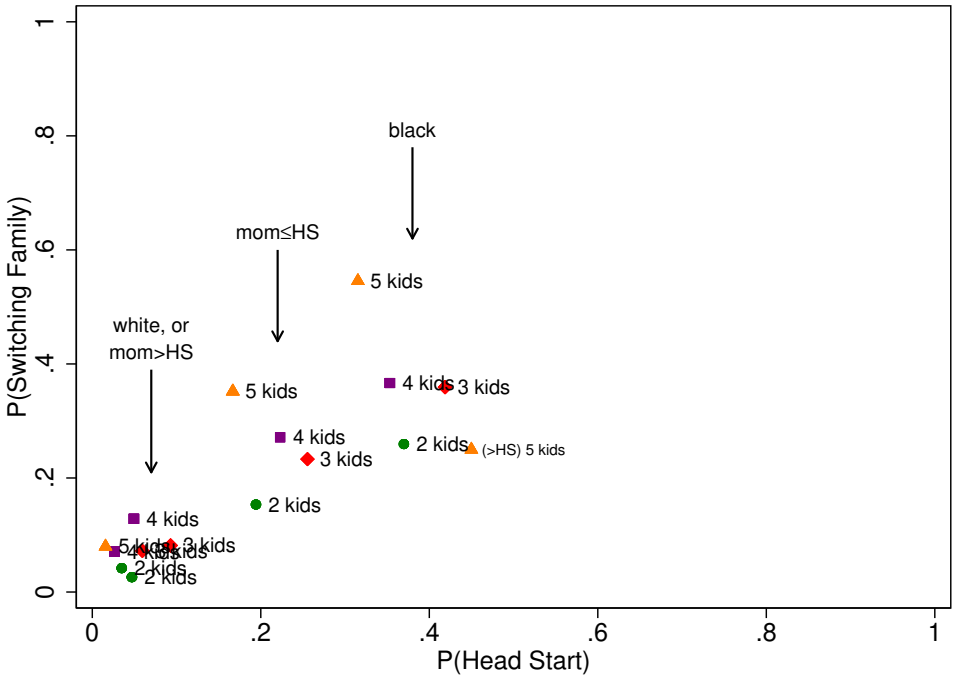
Figure 1: Within-Family Variation in Head Start and Attendance of Some College (PSID)



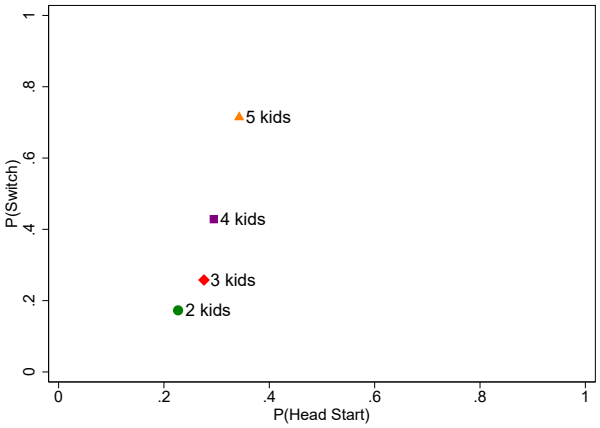
Notes: This figure depicts the identifying variation used in a FFE regression of some college on an indicator for participation in Head Start. Each marker represents the number of individuals that exhibit a particular deviation from the mean Head Start attendance of their family and from the mean attendance 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)

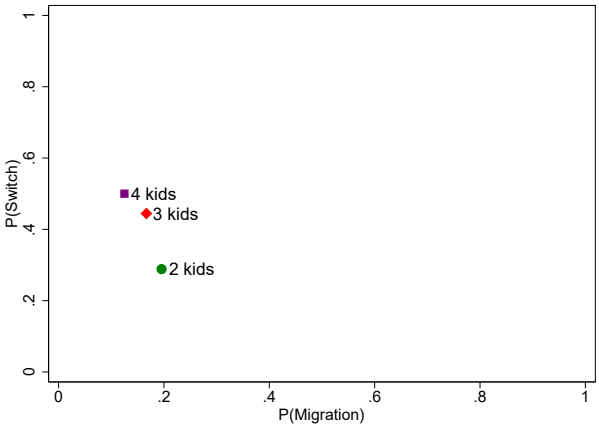
(a) Head Start in PSID



(b) Head Start in CNLSY



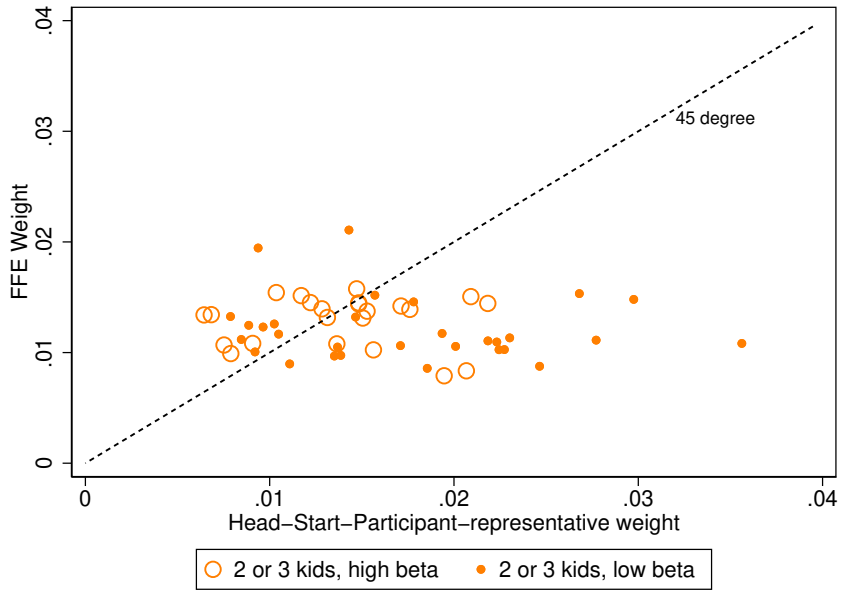
(c) Migration in Census



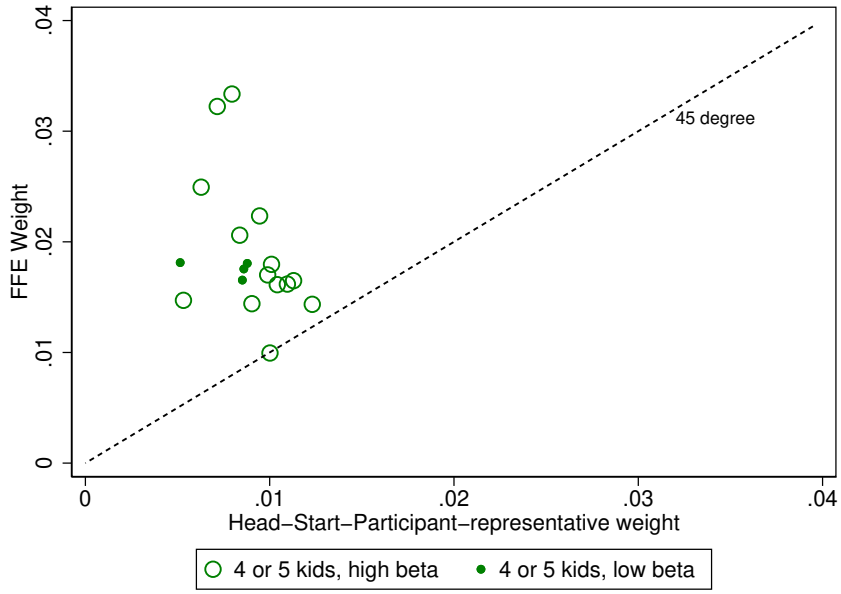
Notes: This figure shows the probability of being in a switching family and the probability of "treatment" by family size using three datasets and varying treatments. Panel (a) plots the probability of being in a switching family and of attending Head Start by family size for the following groups in 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) is a simplified version of (a) using data on Head Start participation and family size from the CNLSY. Figure (c) shows the probability of being in a switching family and the probability of migrating to the northern US, using a linking of the 1910 to 1930 censuses used in Collins and Wanamaker (2014).

Figure 3: FFE Weights and Head-Start-Participant-Representative Weights by Family Size and Some College β (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. This gives a measure of how aligned the characteristics 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. Pr(switch) and Pr(Head Start) are estimated from a multinomial logit model of these outcomes on family size and other covariates 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 ω_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, ω_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_{ig}				
Switchers	80	-0.3	-0.2	1.03
Siblings	80	-0.3	-0.5	1.19
All	80	-0.3	-0.5	1.20
HS Participants	80	-0.3	-0.3	1.04
B. Large family TE; p-score: large family				
Switchers	83.0	-11.1*	-0.6	0.92
Siblings	49.6	22.2*	-0.1	0.70
All	40.3	31.6*	0.1	0.54
HS Participants	41.1	30.7*	0.1	0.55
C. TE linear in X_{ig}; p-score: X_{ig}				
Switchers	94.2	-2.0*	-0.6	1.03
Siblings	80.1	12.2*	1.6*	0.99
All	80.0	12.2*	1.7*	1.00
HS Participants	91.5	0.8	-0.2	1.03
D. TE linear in X_{ig}; p-score: X_{ig} spline				
Switchers	94.2	-1.5*	-0.3	1.04
Siblings	80.1	12.7*	-0.4	1.08
All	80.0	12.8*	-0.4	1.09
HS Participants	91.5	1.3	-0.2	1.09

Notes: This table shows the results from 3,000 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 (using the variable(s) indicated in each of the panel headings as predictors in the multinomial logit step) 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 Reweighting

	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**	0.068	0.026	0.079	0.094**
	(0.108)	(0.053)	(0.060)	(0.062)	(0.056)	(0.042)
Y Mean in Target	–	0.556	0.387	0.437	0.556	
B. Economic Sufficiency Index, Age 30 (PSID)						
Head Start	–	-0.023	-0.038	-0.032	0.021	0.009
	–	(0.102)	(0.086)	(0.098)	(0.088)	(0.090)
Y Mean in Target	–	0.213	-0.198	-0.485	0.213	
C. High School Graduation (CNLSY)						
Head Start	0.086***	0.085***	0.033	0.048	0.020	0.037*
	(0.031)	(0.030)	(0.034)	(0.031)	(0.036)	(0.023)
Y Mean in Target	–	0.776	0.734	0.766	0.776	
D. Idle (CNLSY)						
Head Start	-0.071*	-0.072*	-0.061	-0.055	-0.067	-0.017
	(0.038)	(0.037)	(0.040)	(0.037)	(0.043)	(0.026)
Y Mean in Target	–	0.197	0.221	0.201	0.197	
E. Learning Disability (CNLSY)						
Head Start	-0.059***	-0.059***	-0.031	-0.042**	-0.040**	0.017
	(0.020)	(0.021)	(0.021)	(0.018)	(0.020)	(0.015)
Y Mean in Target	–	0.051	0.055	0.041	0.051	
F. Poor Health (CNLSY)						
Head Start	-0.070***	-0.069***	-0.063**	-0.067**	-0.050*	-0.003
	(0.026)	(0.026)	(0.030)	(0.028)	(0.030)	(0.020)
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 three target populations (shown in the column header) using the post-regression reweighting procedure, in which we multiply group-level estimates of the impact of Head Start by the representative weight for the target population of interest. 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.

Appendices for Selection into Identification in Fixed Effects Models, with Application to head Start

Douglas L Miller, Na'ama Shenhav, and Michel Z. Grosz

INCLUDED WITH MAIN NBER WORKING PAPER

Appendix A: Derivations and Proofs

Appendix B: Supplementary Figures and tables

INCLUDED IN ONLINE APPENDIX FOR NBER WORKING PAPER

Appendix C: Supplementary PSID FFE Results

Appendix D: Replication of GTC (2002)

Appendix E: Functional form choices with Binary Treatment and Binary Outcome

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 $w^t_{g(i)} := \frac{Q_x}{P_x}\frac{Pr(S=1)}{Pr(T=1)}$.

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 $w^t_{g(i)}$ 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.1.1 Extrapolating treatment effects to never-switchers

Note that the ATE for the target population can be written as a weighted average of the ATE for switchers, $\delta^{t,P_x>0}$, and the ATE for never-switchers, $\delta^{t,P_x=0}$:

$$\delta^t := Pr(P_x > 0|T_g = 1) \cdot \delta^{t,P_x>0} + (1 - Pr(P_x > 0|T_g = 1)) \cdot \delta^{t,P_x=0} \tag{22}$$

where $\delta^{t,P_x>0} = \mathbb{E}[Y(1) - Y(0)|T=1, P_x > 0]$ and $\delta^{t,P_x=0} = \mathbb{E}[Y(1) - Y(0)|T=1, P_x = 0]$.

Since we can not identify treatment effects for $P_x = 0$ from the FFE design, we must impose an additional assumption that allows extrapolation of treatment effects for this group.

Assumption 5 (Treatment Effect Functional Form):

$$\mathbb{E}[Y(1) - Y(0)|\mathbf{X_g}] = H(\mathbf{\Phi}; \mathbf{X_g})$$

with $H(\cdot)$ known, and $\mathbf{\Phi}$ parameters that can be consistently estimated. Under Assumption 5, $\hat{\mathbf{\Phi}}$ can be estimated using e.g. the regression $\hat{\delta}_g = \mathbf{\Phi'X_g} + \mathrm{u_g}$.

Proposition 2. *Under Assumptions 1, 2, 3, and 5, and assuming $Pr(P_x > 0|T_g = 1)$ can be consistently estimated, the ATE for the target t can be consistently estimated by*

$$\hat{\delta^t} = Pr(P_x > 0|T_g = 1) \cdot \widehat{\delta^{t,P_x>0}} + (1 - Pr(P_x > 0|T_g = 1)) \cdot \widehat{\delta^{t,P_x=0}}$$

where $\delta^{t,P_x>0}$ comes from Equation 12 and $\delta^{t,P_x=0}$ is estimated from a projection of $\hat{\delta_g}$ on X_g.

Proof:

Define $\widehat{\delta^{t,P_x=0}} := \frac{1}{\sum \mathbf{1}(T_g=1, P_x=0)} \sum_{T_g=1, P_x=0} H\left(\hat{\mathbf{\Phi}}; \mathbf{X_g}\right)$.
Assumptions 1, 3, and 5 imply that

$$plim\ \widehat{\delta^{t,P_x=0}} = \mathbb{E}[Y(1) - Y(0)|T=1, P_x = 0] \tag{23}$$

From the proof of Proposition 1, we have that

$$plim\ \widehat{\delta^{t,P_x>0}} = \mathbb{E}[Y(1) - Y(0)|T=1, P_x > 0] \tag{24}$$

Then,

$$\begin{aligned} plim\ \widehat{\delta^t} &= Pr(P_x > 0|T_g = 1) \cdot \mathbb{E}[Y(1) - Y(0)|T=1, P_x > 0] \qquad (25) \\ &\quad + (1 - Pr(P_x > 0|T_g = 1)) \cdot \mathbb{E}[Y(1) - Y(0)|T=1, P_x = 0] \\ &= \mathbb{E}[Y(1) - Y(0)|T=1] \qquad (26) \end{aligned}$$

A speculative alternative approach would be to take a "double robust" estimation approach. This is modeled after the double robust approach for estimating causal effects, as in Robins, Rotnitzky and Zhao (1995). The discussion in Chapter 17 of Imbens and Rubin (2015, pp. 399-400)

notes that in the traditional setting, consistency of the estimated treatment effect requires either correct specification of the propensity score, or of the regression model. In our setting, instead of trying to model potential outcomes, we are estimating average treatment effects. Implementation would proceed by (i) defining weights $w^t_{g(i)}$ as in equation (13); (ii) estimating $\hat{\delta}_g = H\left(\hat{\mathbf{\Phi}}; \mathbf{X_g}\right)$ using these weights, and then predicting $\frac{1}{\sum \mathbf{1}(T_g=1)} \sum_{T_g=1} H\left(\hat{\mathbf{\Phi}}; \mathbf{X_g}\right)$ over the full target population.

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 the 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 $0.75 \cdot \frac{3}{8}$, and the second family should be given a weight of $0.75 \cdot \frac{5}{8}$.

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:

$$s_{gx} = \frac{\sum(\mathbf{1}(x_i = x)|T_{g(i)} = 1)}{\sum T_{g(i)}} \cdot \frac{\sum \mathbf{1}(g(i) = g)}{\sum(S_{g(i)}|x_i = x)} \tag{27}$$

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,

$$s_{gx} = Pr(x_i = x|T_{g(i)} = 1) \cdot \frac{\sum \mathbf{1}(g(i) = g)}{Pr(x_i = x|S_{g(i)} = 1) \cdot \sum S_{g(i)}} \tag{28}$$

$$= \frac{Pr(x_i = x|T_{g(i)} = 1)}{Pr(x_i = x|S_{g(i)} = 1)} \cdot Pr(g(i) = g|S_{g(i)} = 1) \tag{29}$$

A.2.3 Estimation

1. We obtain an estimate of $\hat{Q}_x = Pr(T_{g(i)} = 1|x_i = x)$ as fitted values from a regression of T on X.

   This is equal to $\frac{Pr(x_i=x|T_{g(i)}=1) \cdot Pr(T_{g(i)}=1)}{Pr(x_i=x)}$ 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 $\frac{Pr(x_i=x|T_{g(i)}=1)}{Pr(x_i=x|S_{g(i)}=1)}\cdot\frac{Pr(T_{g(i)}=1)}{Pr(S_{g(i)}=1)}$.

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}. 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 individual level: $w_i^t = \frac{Q_{X,C}}{P_{X,C}} \cdot \frac{Pr(S=1)}{Pr(T=1)}$.

The assumptions from earlier now must now be modified slightly to recover the ATE. Assumption 1, the conditional fixed effects assumption now requires $\epsilon_{ig}, u_{ig} \perp\!\!\!\perp D_i|C_i, \alpha_g$. This gives that $E\left[\hat{\delta}_g\right] = \delta_g$.

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

This model can be adapted to allow treatment effects to vary with individually-varying covariates, such as gender of the treated individual. One approach is to re-write the treatment effects as a function of a group-level measure of these covariates (e.g. the share of Head Start participants in the group that are female). Then our main estimaton framework applies. These group-level measures are included in the prediction of P_x and Q_x which should be sufficient to recover the ATE.

An alternative approach is to add a $\beta_C \cdot (C_i D_i)$ term to the estimating equation. Our principal re-weighting method can be used to estimate the average of the δ_g for the target population; and this average could then be added to $\hat{\beta}_C$ times the average covariate C_i value for the target population to arrive at an estimate for the ATE for this group.

A.3.2 Defining residual switchers

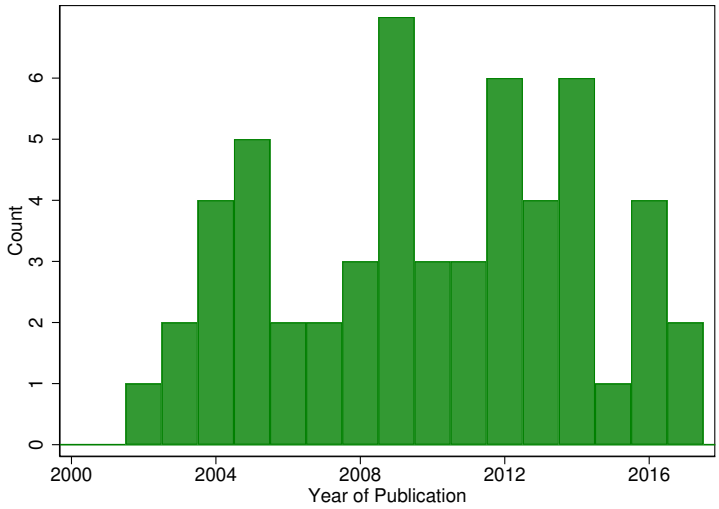
Consider the "deviations from group means" projection matrix, $M = I - H\left(H'H\right)^{-1}H$, with H a matrix of dummy variables for group membership: $H\left[i,j\right] = 1$ if unit i is a member of group j, and 0 otherwise. Let $\widetilde{D} = M \cdot D$ be deviations in treatment from group means. 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.

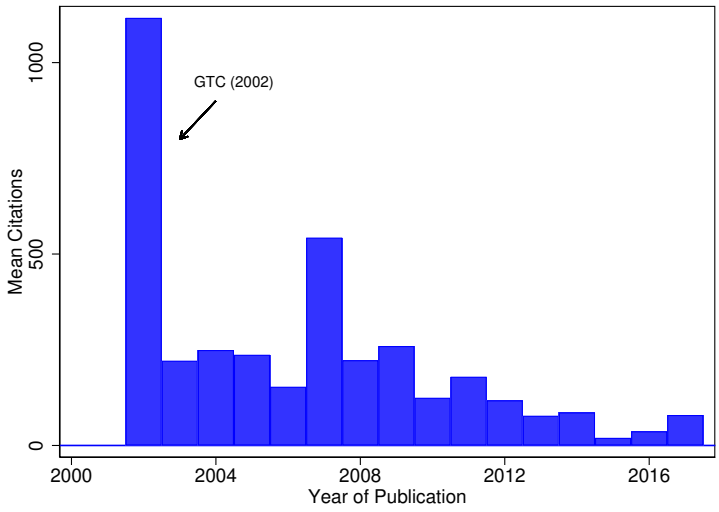
B Supplementary Figures and Tables

Figure B.1: Popularity of Family Fixed Effects Articles

(a) Publications by Year



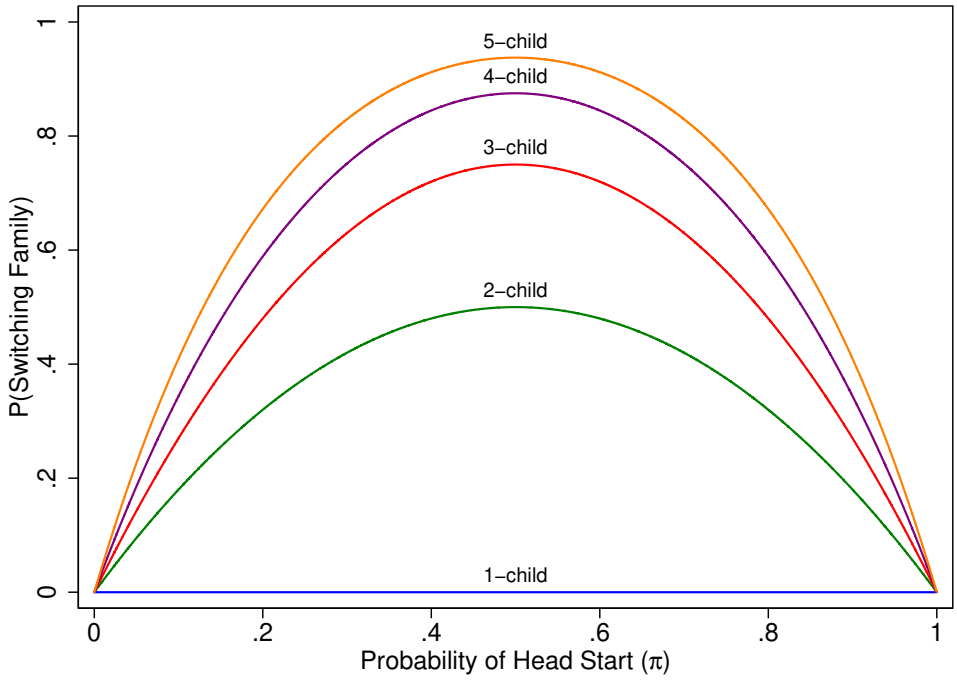
(b) Average Citations by Year of Publication



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.

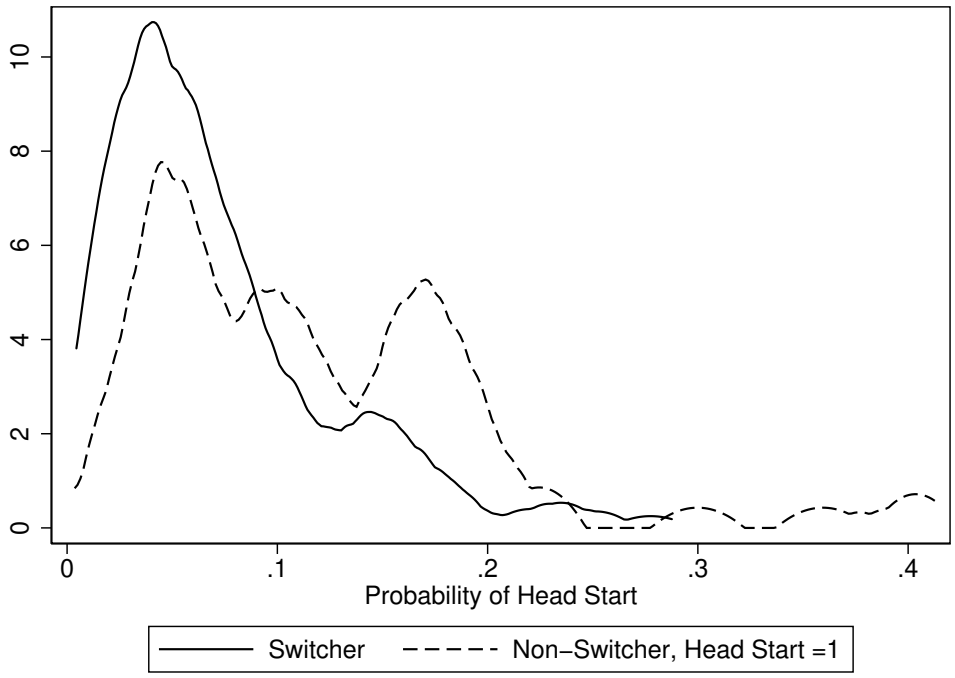
Figure B.2: Illustrative Model of the Role of Family Size in Switching

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



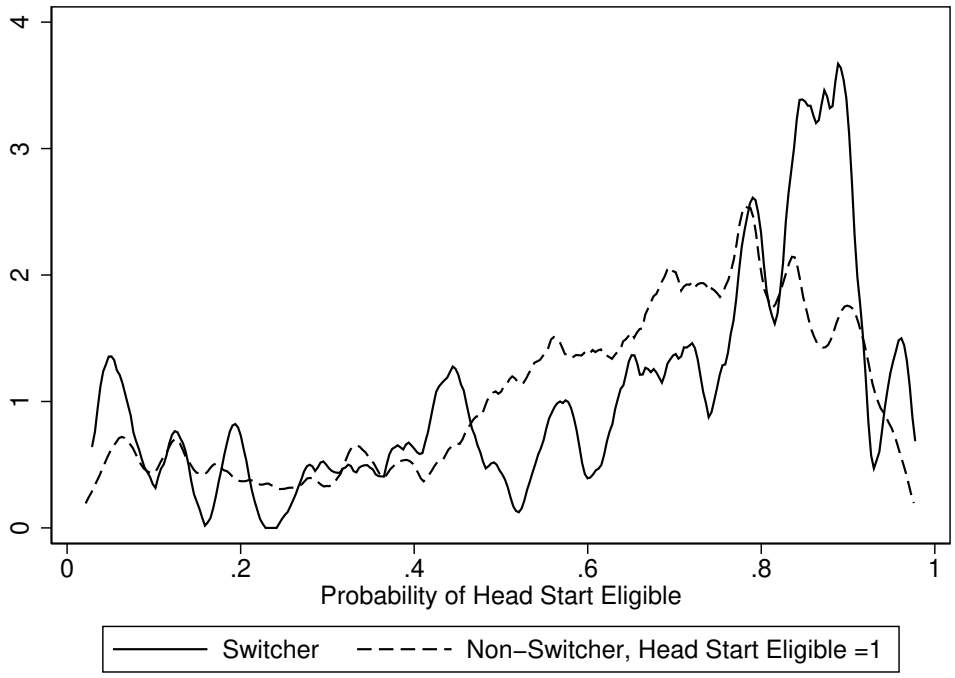
Notes: This figure plots the theoretical function: $P(HSSwitchingFamily) = 1 - (1 - \pi)^{n_g} - \pi^{n_g}$, where n_g is the number of children in a family and π is the probability of attending Head Start, for 2-, 3-, 4-, and 5 (plus)- child families.

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. The are 4 non-switchers (5%) who have a probability of being in Head Start that is outside of the support of the switcher sample.

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 (π)	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 Ignorability Assumption:

Do Individuals in the Target Population Have Differential Treatment Effects?

	Target= Eligible	Target= Participants
Some College (Whites, PSID)		
In Target	–	-0.084
	–	(0.124)
Observations	–	213
Economic Sufficiency Index (Whites, PSID)		
In Target	–	0.145
	–	(0.199)
Observations	–	213
High School Graduation (NLSY)		
In Target	0.014	-0.101
	(0.069)	(0.062)
Observations	467	581
Idle (NLSY)		
In Target	0.036	0.033
	(0.085)	(0.097)
Observations	467	581
Learning Disability (NLSY)		
In Target	-0.025	-0.095
	(0.046)	(0.063)
Observations	467	581
Poor Health (NLSY)		
In Target	0.015	-0.026
	(0.068)	(0.058)
Observations	467	581

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 for balance on observables: target individuals are assigned a weight of $Pr(T_g = 1, S_g = 1|X_{ig})$ and non-target individuals are assigned a weight of $Pr(T_g = 0, S_g = 1|X_{ig})$. The first two panels use data from the PSID white sample, and the final four panels use data from the CNLSY. There are no individuals in the PSID white switcher sample that are ineligible for Head Start, which causes us to have missing estimates ("–").

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	-0.034	-0.030	-0.029
	(0.098)	(0.045)	(0.043)	(0.049)	(0.050)
Y Mean in Target	–	0.921	0.852	0.848	0.921
B. Good Health Index, Age 30					
Head Start	–	-0.265	-0.253	-0.439	-0.162
	–	(0.249)	(0.263)	(0.313)	(0.316)
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	-0.017	-0.014	-0.015
	(0.065)	(0.031)	(0.025)	(0.026)	(0.024)
Y Mean in Target	–	0.862	0.854	0.896	0.862
B. Some College					
Head Start	0.023	-0.016	-0.031	-0.028	-0.032
	(0.066)	(0.036)	(0.032)	(0.034)	(0.032)
Y Mean in Target	–	0.396	0.376	0.423	0.396
C. Economic Sufficiency Index, Age 30					
Head Start	–	-0.023	-0.190**	-0.211***	-0.167**
	–	(0.102)	(0.072)	(0.073)	(0.071)
Y Mean in Target	–	-0.552	-0.626	-0.674	-0.552
D. Good Health Index, Age 30					
Head Start	–	-0.265	0.052	0.062	0.033
	–	(0.249)	(0.146)	(0.161)	(0.133)
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: Head Start Impact for Representative Eligible Children, Participants, and Siblings Using Reweighting with Regression Extrapolation to Singletons

	Reweighted ATE, Target =	
	HS Eligible	Participants
A. Some College (PSID)		
Head Start	0.083	0.009
	(0.102)	(0.109)
Y Mean in Target	0.387	0.437
B. Economic Sufficiency Index, Age 30 (PSID)		
Head Start	-0.056	0.003
	(0.139)	(0.229)
Y Mean in Target	-0.198	-0.485
C. High School Graduation (CNLSY)		
Head Start	0.041	0.047
	(0.032)	(0.030)
Y Mean in Target	0.734	0.766
D. Idle (CNLSY)		
Head Start	-0.059	-0.057
	(0.038)	(0.037)
Y Mean in Target	0.221	0.201
E. Learning Disability (CNLSY)		
Head Start	-0.034*	-0.041**
	(0.020)	(0.017)
Y Mean in Target	0.055	0.041
F. Poor Health (CNLSY)		
Head Start	-0.059*	-0.065**
	(0.028)	(0.030)
Y Mean in Target	0.098	0.074

Notes: Columns 1 and 2 present reweighted estimates of the effect of Head Start for the Head-Start-eligible and Head-Start-participant target populations, where the treatment effects for singletons are extrapolated from switchers using OLS. Sample size is N=2,986 for the expanded sample, and 1,036 for GTC. Standard errors obtained using bootstrap. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table B.13: FFE Estimates Reweighted using Gibbons, Suarez, Urbancic (2018) Method

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

Notes: Column 1 reprints the FFE estimate using our expanded sample for PSID outcomes and using our replication sample for CNLSY outcomes. 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. 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.14: 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.