



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



Douglas L. Miller
Na'ama Shenhav
Michel Grosz

ABSTRACT

Many papers use fixed effects (FE) to identify causal impacts. We document that when treatment status only varies within some FE groups (for example, families, for family fixed effects), FE can induce nonrandom selection of groups into the identifying sample. To address this, we introduce a reweighting-on-observables estimator that can help recover the average treatment effect for policy-relevant populations. We apply these insights to reexamine the long-term effects of Head Start in the PSID and the CNLSY and find that the reweighted estimates are frequently smaller than the FE estimates. This underscores concerns with the external validity of FE estimates. The tools that we propose can strengthen the validity of this approach.


Doug Miller is at the Brooks School of Public Policy and the Economics Department, Cornell University (dlm336@cornell.edu). Na'ama Shenhav is at the Department of Economics, Dartmouth College (naama.shenhav@dartmouth.edu). Michel Grosz is at the Bureau of Economics, Federal Trade Commission (mgrosz@ftc.gov). The authors thank Colin Cameron, Liz Cascio, Janet Currie, Hilary Hoynes, Pat Kline, Erzo F.P. Luttmer, Jordan Matsudaira, Zhuan Pei, 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. They are grateful to Alex Magnuson, Wenran Li, Wenrui Huang, Jack Mueller, and Mary Yilma for excellent research assistance. The views expressed in this article are not necessarily those of the Federal Trade Commission nor any of its commissioners. The replication files for this article are available online at <https://doi.org/10.5281/zenodo.7761792>.


[Submitted November 2018; accepted May 2021]; doi:10.3368/jhr.58.5.0520-1093OR1

JEL Classification: I38, I28, and C23

ISSN 0022-166X E-ISSN 1548-8004 © 2023 by the Board of Regents of the University of Wisconsin System

 Color version of this article is available online at: <https://jhr.uwpress.org>.

 Supplementary materials are available online at <https://jhr.uwpress.org>.

 This open access article is distributed under the terms of the CC-BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0>) and is freely available online at: <http://jhr.uwpress.org>

Douglas L. Miller <https://orcid.org/0000-0001-5906-7822>

Na'ama Shenhav <https://orcid.org/0009-0007-2910-3539>

Michel Z. Grosz <https://orcid.org/0000-0003-3814-0650>

I. Introduction

Fixed effects (FE) are frequently used to obtain identification of the causal impact of an attribute, intervention, or policy. These models have been used to identify the impacts of academic peers (school–grade FE; Hoxby 2000; Carrell and Hoekstra 2010), criminal peers (facility–offense FE; Bayer, Hjalmarsson, and Pozen 2009), the local healthcare environment (individual FE; Finkelstein, Gentzkow, and Williams 2016), participation in means-tested programs (family FE [FFE]; Currie and Thomas 1995; Garces, Thomas, and Currie 2002; Deming 2009; Rossin-Slater 2013), neighborhood quality (FFE; Chetty and Hendren 2018a), 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 targeting the average treatment effect for a policy-relevant population (for example, participants). However, in contrast with other common estimators, there is not yet a comprehensive framework for considering the *external validity* of FE estimates.

We document that FE can induce a special type of systematic selection in estimation, which we term *selection into identification* (SI). SI occurs because FE estimates are only identified from FE groups (for example, families, for FFE) that have variation in treatment (“switchers”), which may exclude some groups.¹ This is a distinct problem from whether within-group comparisons are internally valid, which has been the typical subject of debate for FE estimators (for example, Bound and Solon 1999), and which is not the focus of this paper.

While researchers may be aware of this issue in an abstract sense, our review of the applied literature indicates that it is rarely discussed and almost never empirically examined. This may be because researchers assume that SI is a minor concern or because there is not a user-friendly tool to address it. We find across multiple previous applications of FE that switchers are typically a small subset of the sample and systematically different from the overall population. This highlights a new and important concern about the external validity of FE estimates.² We also show that reweighting-on-observables methods can be used to help “undo” SI to recover the average treatment effect for policy-relevant populations. We apply these insights to reexamine the external validity of prior FFE estimates of the long-run impact of Head Start, a federally funded preschool program.

We begin by illustrating SI in a well-known application of FE: an FFE model where the treatment of interest is an indicator for whether an individual attended Head Start (see, for example, Garces, Thomas, and Currie 2002). We use the Panel Study of Income Dynamics (PSID), as in Garces, Thomas, and Currie (2002).³ First, we show that this model uses substantially fewer identifying groups relative to an estimation model without FE. Among the 5,355 children who have siblings in our sample, only 1,098 children reside in switcher households. Second, we find that the loss of sample variation is systematically

1. In the presence of control variables that vary within a group, there may be variation among nonswitchers “net of controls.” We focus primarily on cases where this phenomenon is small in magnitude, but include a discussion of this issue in Section V.

2. SI is distinct from issues related to conditional variance weighting in Gibbons, Suárez, and Urbancic (2019), which focuses on cases where all groups are switchers.

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) and Currie and Rajani (2015). For health, see Almond, Chay, and Lee (2005); Figlio et al. (2014); Abrevaya (2006); Black, Devereux, and Salvanes (2007); and Xie, Chou, and Liu (2016), among others. We summarize the prevalence of this design in Section II.

related to observables. In particular, families that have a very low likelihood of attending Head Start or a very high likelihood of attending Head Start are both unlikely to be switchers, while families with more children are significantly more likely to be switchers. As a result, switchers are not representative of the overall sample or Head Start participant families along multiple dimensions. Third, SI varies across subgroups—the FFE identifying sample misses 93 percent of the sibling sample for white children, but only 62 percent of the sample for black children.

The main implication of SI is that under heterogeneous treatment effects, the FE estimate is no longer representative of the ATE for the sample. We show that this bias is not addressed by “undoing” conditional variance weighting among switchers (Gibbons, Suárez, and Urbancic 2019) and is larger in our applications than that from conditional variance weighting.⁴ A second result is that part of the difference between the ordinary least squares (OLS) estimate and FE estimate reflects the disparity in the identifying sample—that is, the causal treatment effect for nonswitchers—and should not be interpreted as solely reflecting OLS bias.⁵

To help recover the average treatment effect (ATE) for policy-relevant populations (“targets”), we develop a method for reweighting FE estimates that builds on extrapolation techniques from the experimental and instrumental variables (IV) literatures.⁶ In particular, we show that the ATE for a given target can be computed by weighting group-level FE estimates by the ratio of two propensity scores: (i) the propensity to be in the target (for example, a program participant) and (ii) the propensity to be in the switcher population. This reweighting is valid under two key assumptions: first, that these propensity scores can be estimated using observable covariates, and second, that treatment effects are uncorrelated with being a switcher conditional on covariates. We discuss testable implications of these assumptions and show empirical support for them in our applications. These assumptions provide a weaker alternative to the typical homogeneous treatment effects assumption required for the FE estimate to deliver the ATE. However, in some cases, a researcher might prefer the less externally valid FE estimate rather than rely on these assumptions.⁷

We demonstrate the performance of our reweighting using Monte Carlo simulations. We find that reweighting reduces or eliminates bias relative to FE in the presence of covariate-based treatment heterogeneity. We also 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.

4. This is consistent with Gibbons, Suárez, and Urbancic (2019), whose findings suggest that the bias from conditional variance weighting is less than 5 percent for most estimates.

5. Although intuitive, estimating similar OLS coefficients across the overall sample and the switcher sample does not guarantee that the FE estimates are generalizable. This is because OLS bias can vary across covariate values that predict switching. In our application, for example, OLS bias varies with family size (see Section III. C), such that the family-size gradient in treatment effects is much stronger for FE estimates than OLS estimates. As a result, a simple comparison of the OLS coefficients across the OLS and switcher samples will understate the impact of having more large families in the switcher sample.

6. See Angrist and Fernandez-Val (2013) and Aronow and Carnegie (2013) for extrapolation from IV and Stuart et al. (2011) and Andrews and Oster (2019) for extrapolation from experiments.

7. In these cases, researchers can note that FE estimates are specific to the population of switchers and characterize that population.

Based on these findings, we propose new standards for practice when presenting results using FE research designs. Researchers should: (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 nonswitcher families (for example, looking ahead to Table 2), and (iii) reweight FFE estimates for a representative population (for example, looking ahead to Table 6). Reweighted estimates can be presented either as an additional diagnostic tool or as a primary measure of treatment effects. We are not the first to use the more rigorous reporting standards in the first two standards, but in our survey of the FFE literature, the vast majority do not discuss either of these issues. Out of 35 papers with binary treatments, only one-third reported the sample sizes we recommend, and just one paper included showing the balance of covariates across switcher and nonswitcher families.⁸

We apply these insights to revisit past 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 percent of the number of three- and four-year-old children in poverty, which makes it a quantitatively important intervention for this population (Carnerio and Ginja 2014).⁹ 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; Garces, Thomas, and Currie 2002), which find positive impacts on economic and noncognitive outcomes of participants measured in adulthood.

First, using data from the PSID and the Children of the National Longitudinal Study of Youth (CNLSY) (as in Garces, Thomas, and Currie 2002; Deming 2009), we newly document that Head Start has greater returns in larger families across multiple human capital measures.¹⁰ 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.¹¹

Second, we show that the FFE estimate of the impact of Head Start is typically too large relative to the reweighted ATE for participants. Using the PSID, we obtain a FFE estimate that suggests that Head Start leads to a statistically significant 12 percentage point increase in attendance of some college for white individuals. However, when we reweight the FFE estimates to obtain the ATE for white Head Start *participants*, we find that Head Start leads to a 2.1 percentage point increase in the likelihood of attending some college (SE = 5.9 percentage points). This reweighted ATE for participants is 83 percent smaller than the FFE estimate ($p < 0.05$ for the difference.) The participants’ ATE is also 93 percent smaller than the estimated effects on college

8. Important exceptions include Finkelstein, Gentzkow, and Williams (2016) and Wiswall (2013), who include a substantive discussion of external validity concerns, and Currie and Rossin-Slater (2013), who show the change in summary statistics when moving from the sample of mothers with multiple births to the sample that had at least one hurricane during pregnancy, and are therefore likely to be “switchers.” Garces, Thomas, and Currie (2002) 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 preschool, Head Start, and no-formal-care coefficients, but not for each coefficient.

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

10. Although we rely on the PSID, as in Garces, Thomas, and Currie (2002), we use a sample of siblings three times as large as in Garces, Thomas, and Currie (2002) due to the additional data that has been collected since the publication of that paper and somewhat different inclusion criteria.

11. In Section VI, 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.

attendance in Garces, Thomas, and Currie (2002), 45 percent to 91 percent smaller than unadjusted estimates for all participants from other FFE studies (Bauer and Schanzenbach 2016; Deming 2009), and 51 percent smaller than estimates from the county rollout of Head Start (Bailey, Sun, and Timpe 2021).

Similarly, in the CNLSY, the FFE estimate suggests that Head Start leads to an 8.5 percentage point increase in high school completion, while the reweighted estimate for Head Start participants is 44 percent smaller and not statistically significant ($p < 0.10$ for the difference). Reweighting also attenuates the previously estimated impact of Head Start on idleness and having a learning disability, relative to the FFE estimates. In contrast, reweighting has little to no impact on the estimates for the poor health outcome.

Finally, we discuss the broader relevance of SI for panel fixed effects models, particularly for short panels and “lumpy” treatment variables (for example, binary treatments). In particular, we walk through how the guidelines that we have put forth could apply to three prior FE applications.

This work makes three main contributions. First, we empirically document across multiple FE settings that many groups are not switchers and that switchers are observably different than nonswitchers.¹² This stands in contrast to the assumption that all FE groups are switchers, which underlies several econometric techniques (Bates et al. 2014; Loken, Mogstad, and Wiswall 2012; Gibbons, Suárez, and Urbancic 2019). It also implies that the method in Gibbons, Suárez, and Urbancic (2019) for recovering the ATE from FE models by “undoing” the conditional variance weighting—which relies on this assumption—is not sufficient to recover the ATE in many contexts. Finally, it suggests that more researchers using FE models should examine SI and discuss this in their interpretation of FE estimates. We show that this is rarely done in current practice and provide easy rules-of-thumb to guide such investigations.

Second, we develop a general reweighting method to obtain the ATE for policy-relevant populations from FE models.¹³ This provides an alternative to defining the population of interest to be switchers, as discussed in Angrist (1998) (for OLS models) and Imai and Kim (2019) (for FE models). In doing so, we add to a small set of papers that provide methods for researchers to extrapolate FE treatment effects beyond switchers. Existing approaches include bounding methods for nonparametric panel models (Section 4 of Chernozhukov et al. 2013) and using a random coefficients model for treatment effect heterogeneity (Wooldridge 2005, 2019; Bates et al. 2014). Our literature review suggests that practitioners do not pursue either of these current methods for correction. Relative to these other correction methods, we provide easy-to-implement guidelines for using our reweighting approach that is closely fitted to current empirical practice.

Our reweighting solution is most similar in spirit to Angrist and Fernandez-Val (2013), who reweight IV estimates using discrete covariates.¹⁴ We tailor this idea to the

12. This is distinct 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. See, for example, Goodman-Bacon (2018); Borusyak and Jaravel (2017); Callaway and Sant’Anna (2018); and Chaisemartin and D’Haultfoeille (2019).

13. This is distinct from strategies that use reweighting for internal validity, such as traditional propensity score estimation methods.

14. More broadly, we relate to a number of studies that use reweighting to account for the discrepancy between “what you want” and “what you get” from common estimators. These include Lochner and Moretti (2015), who reweight OLS with IV weights for greater comparability; Słoczyński (2018), who reweights OLS to obtain the

FE context to develop our reweighting method, and we show the set of assumptions under which these weights can recover the ATE from FE estimates. Also different from Angrist and Fernandez-Val (2013), we reweight using propensity scores, which allow for greater flexibility in conditioning variables.

Third, our empirical findings 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 2018; Bauer and Schanzenbach 2016; Johnson and Jackson 2019; Bailey, Sun, and Timpe 2021; Pages et al. 2020; Barr and Gibbs 2022; De Haan and Leuven 2020). in addition to the FFE papers above). These studies typically present local average treatment effect (LATE) or intention-to-treat estimates and find improvements in childhood health, reductions in adolescent behavioral problems and obesity, and increases in adult educational attainment and earnings.¹⁵ 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 reweighting to get closer to the ATE for Head Start participants.

II. A Survey of Family Fixed Effects Applications

We begin by performing a survey of FE papers in the literature to gauge the prevalence of these methods. 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 that 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 or secondary strategy.¹⁶

Our literature review yields 55 papers published during 2002–2017. We provide descriptive statistics of these articles in Table 1, including statistics by journal. Overall, these articles account for between 0.5 and 1 percent of the papers published in our sample of journals in any given year, but this varies from 0 to 9 percent of each journal in a given year. 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

ATE; and Stuart et al. (2011) and Andrews and Oster (2019), who propose reweighting experiments to account for selection into participation.

15. One exception to this is Pages et al. (2020), who suggest that the effect of Head Start may be negative for recent cohorts, although the identifying sample is not discussed.

16. 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*, and *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 that 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.

Table 1*Family Fixed Effects Articles in Top Applied Journals 2002–2017*

	Binary Indep.	Binary Dep.	Both Binary	Total
<i>AEJ: Applied</i>	6	4	3	8
<i>AEJ: Economic Policy</i>	1	1	1	1
<i>AER</i>	3	1	1	5
<i>AER Papers and Proceedings</i>	2	2	1	3
<i>Journal of Health Economics</i>	5	3	2	7
<i>Journal of Human Resources</i>	7	2	2	12
<i>Journal of Labor Economics</i>	2	1	1	5
<i>Journal of Political Economy</i>	2	1	1	2
<i>Journal of Public Economics</i>	4	4	4	5
<i>QJE</i>	1	4	1	4
<i>Review of Economics and Statistics</i>	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 <i>N</i>	Total <i>N</i>		
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 indep.	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 identified 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.

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 percent 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 percent of papers have fewer than 1,200 observations, while the top 25 percent have more than 160,000 sibling observations.

Online 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, for example, Chetty and Hendren (2018a,b). We discuss the prevalence of switcher counts in the next section.

III. Fixed Effects and Selection into Identification

We use an FE research design in our application 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 have siblings participate in Head Start is influenced by low parental income (a requirement for eligibility), which may independently influence long-term outcomes. As a result, 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 (for example, Head Start) and $g(i)$ be the relevant group (for example, 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. 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 group as simply g . We refer to groups for whom realized $\text{Var}[D_i | i \in g(i)] > 0$ as “switchers,” and we 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, for example, mean family income, but is randomly assigned within groups:

Assumption 1: Group ID Conditional Independence

$$(1) \quad Y_i(0), Y_i(1) \perp D_i | g(i) = g$$

Assumption 1 encompasses the standard FE specification assumption 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, $Y_i(1) - Y_i(0)$.¹⁷ In the

17. 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).

context of Head Start, within-family variation in treatment has been shown to be uncorrelated with most observable characteristics of children (Deming 2009; Garces, Thomas, and Currie 2002), suggesting the assumption is reasonable.

Under this assumption, estimated treatment effects $\hat{\delta}_{g,FE}$ are an unbiased (but very noisy) estimate of group-level treatment effects, $\delta_g \equiv \mathbb{E}[Y_i(1) - Y_i(0) | g(i) = g]$. The FE estimate averages $\hat{\delta}_{g,FE}$ 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 , n_g (Angrist 1998; Angrist and Pischke 2009, their Equation 3.3.7). The FE estimand is:

$$(2) \quad \delta_{FE} = E_g [\hat{\delta}_g \cdot \omega_{g,FE}]$$

where

$$\omega_{g,FE} = \frac{\text{Var}[D_i | g(i) = g, S_g = 1] \cdot n_g \cdot \text{Pr}[S_g = 1 | g(i) = g]}{\sum_g \{ \text{Var}[D_i | g(i) = g, S_g = 1] \cdot n_g \cdot \text{Pr}[S_g = 1 | g(i) = g] \}}$$

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 due to the fact that $\text{Pr}[S_g = 1 | g(i) = g]$ is not likely to be the same for all groups. The first issue is well understood in principle, but the degree to which G_S is smaller than G is often underappreciated, and implicitly assumed to be negligible in many theoretical results. The second issue is more novel and should cause researchers to update the interpretation of the population for which these estimates are relevant. In our literature review, both of these issues are not reported in empirical practice: one-third of papers with a binary independent variable reported the number of switchers, and one paper showed summary statistics by switching status.¹⁸

A. Empirical Relevance

To illustrate these two methodological issues, we use the data from our empirical application, which is described in detail in Section VI. The sample consists of 2,986 white children born in the years 1954–1987. The regression of interest 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 (SE=0.044). When FFE are added, the coefficient becomes 0.120 (SE=0.053). This 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 their family,

18. We also note that the goal of the latter comparison was to show that within-family comparisons were more valid than cross-sectional comparisons, not to discuss external validity. Even if we look at the more recent period after 2010, just five out of 20 articles reported the number of switchers, suggesting that researchers still do this rarely.

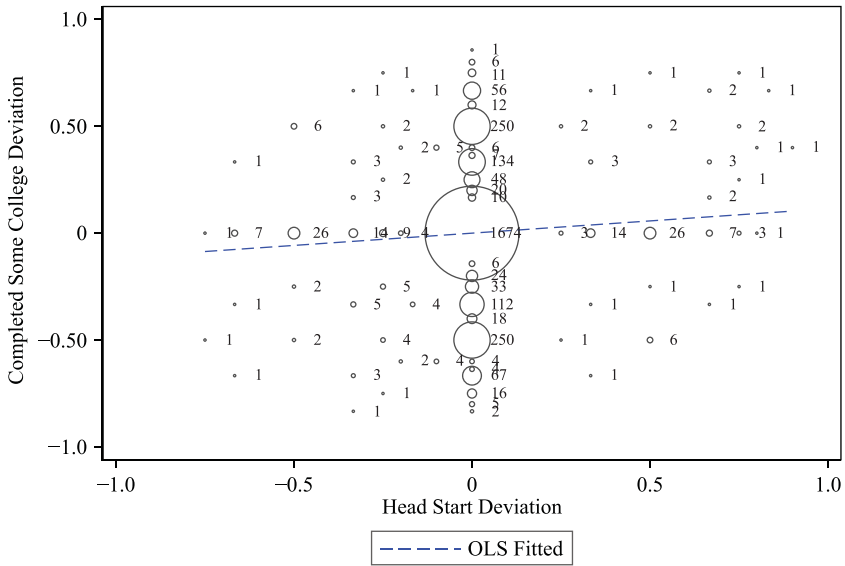


Figure 1
Within-Family Variation in Head Start and Attendance of Some College (PSID)

Source: Panel Study of Income Dynamics 1968–2011 waves.

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.

$g(i), \text{HeadStart}_i - \overline{\text{HeadStart}_{g(i)}}$, against the within-family deviation in attainment of some college for the sample, $\text{AnyCollege}_i - \overline{\text{AnyCollege}_{g(i)}}$.¹⁹ 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 percent of the sample, just 213 individuals are in switching families.

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

19. 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. For example, 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.

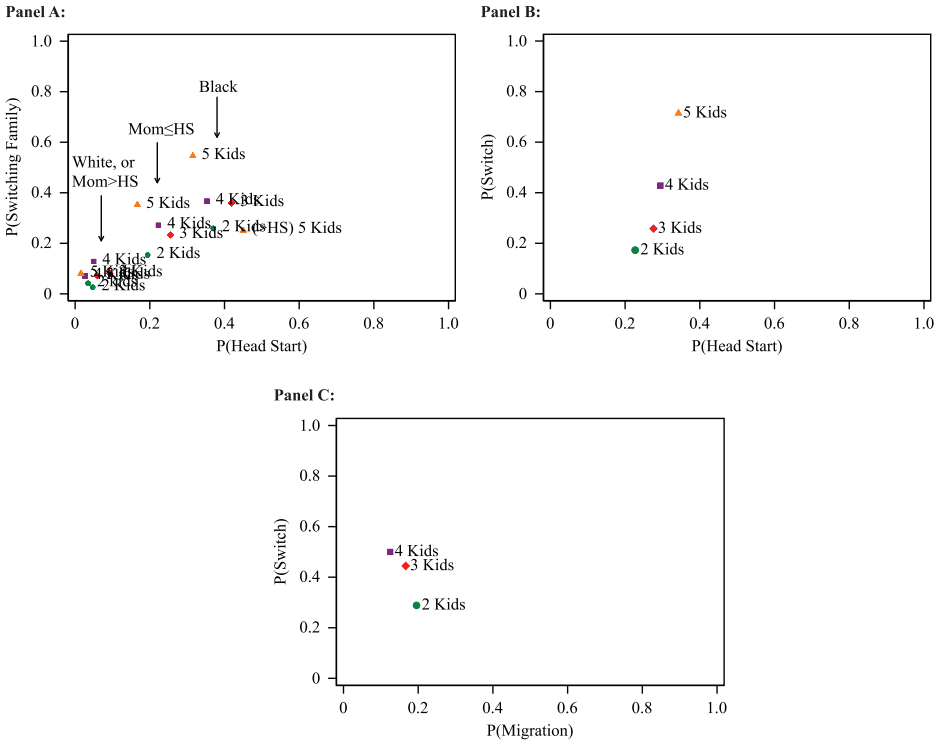


Figure 2
Likelihood of Being a Switcher Family Increases with Family Size and Probability of Treatment

Notes: This figure shows the probability of being in a switching family and the probability of “treatment” by family size using three data sets 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. Panel B is a simplified version of Panel A, using data on Head Start participation and family size from the CNLSY. Panel C shows the probability of being in a switching family and the probability of migrating to the northern United States, using a linking of the 1910–1930 censuses used in Collins and Wanamaker (2014).

family, then the probability of switching, $Pr(S_g = 1)$ 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 [Online 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.²⁰ 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 sample used for FFE identification is comprised of 7 percent of the sibling sample for whites and 38 percent 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 socioeconomic status, that determines π .²¹

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.

To characterize the effects of SI more broadly, we examine how switcher- and non-switcher families compare across a large number of observable characteristics. Panel A of Table 2 indicates that in addition to having more siblings, children in switcher families tend to have less-educated parents (Column 3). These differences in parental education are significant (at the 10 percent level) even in a regression framework where we control for differences in family size and the other covariates in the table (Columns 4 and 5). Family income during preschool is also significantly lower in switcher families.²²

Next, we examine a one-dimensional summary of how much overlap there is in the characteristics of switchers and nonswitchers. We do so by comparing propensity score–type weights across switchers and nonswitchers. This is in the spirit of the literature on using propensity score to extrapolate experimental results to broader populations (Stuart et al.

2011). For comparison we use propensity-score-derived weights, $\frac{Pr(S_{g(i)} = 1 | \mathbf{X}_{ig})}{Pr(\text{HeadStart}_i = 1 | \mathbf{X}_{ig})}$,

which we discuss in more detail in Section IV. These weights give 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 one implies perfect

20. [Online 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 percent of two-child families have no Head Start participants, compared with 48 percent of families with five or more children.

21. In [Online Appendix Figure B.3](#), we create a more detailed version of this figure by graphing the share switching by family size against the predicted probability of attending Head Start. It shows that the level of switching is lower than in the theoretical model, suggesting that decisions to attend Head Start are not truly independent across children. However, the general metaphor holds: switching increases with $\text{pr}(\text{Head Start})$ and with family size, even if the magnitude is smaller.

22. If we limit ourselves to families with Head Start participants, we continue to find many statistically significant differences (see [Online Appendix Table B.2](#)).

Table 2
Switchers and Nonswitchers Vary along Dimensions Other than Family Size

	Switch (1)	Nonswitch (2)	<i>p</i> -Value (1)=(2) (3)	Beta Switch (4)	<i>p</i> -Value (4) (5)
Panel A: Individual Covariates					
Fraction female	0.562	0.495	0.001	0.024	0.472
Fraction African-American	0.516	0.111	0.000	0.249	0.000
Mother's years education	9.283	11.230	0.000	-0.140	0.453
Father's years education	9.190	11.371	0.000	-0.389	0.075
Had a single mother at age 4	0.252	0.099	0.000	0.055	0.011
Family income (age 3–6) (CPI adjusted)	31809	52574	0.000	-4759	0.000
Mother employed, age 0	0.508	0.570	0.013	0.055	0.019
Mother employed, age 1	0.517	0.543	0.281	0.058	0.018
Mother employed, age 2	0.536	0.554	0.439	0.118	0.000
Household size at age 4	5.487	4.451	0.000	0.755	0.000
Fraction low birth weight	0.077	0.058	0.075	0.010	0.483
Observations	1103	5500	6603	7372	7372
Panel B: Inverse Selection into Identification Weights					
Pr(switch)/Pr(Head Start), Whites	2.993 (2.17)	2.347 (1.95)			
Pr(switch)/Pr(Head Start), Blacks	1.969 (1.29)	1.137 (1.03)			

Source: Panel Study of Income Dynamics 1968–2011 waves.

Notes: Panel A of this table presents comparisons of the characteristics of individuals in switching families and nonswitching 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 nonswitcher families. Column 3 presents the *p*-value 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 *p*-value shown in Column 5, with standard errors clustered on id1968. All controls from the main specification are included, except 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 parentheses) 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 one implies perfect alignment, while a higher value implies that the characteristics of switchers are overrepresented 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.

alignment, while a higher value implies that the characteristics of switchers are overrepresented 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 3 for the switchers sample, which is at least 0.6 SD larger than for nonswitchers. This indicates that the

observables of switchers are not aligned with the population of interest and that this misalignment is worse for switchers than nonswitchers.²³

B. 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 four-sibling family with one treated and three untreated individuals has an $\omega_{g,FE}$ that is 25 percent smaller than the $\omega_{g,FE}$ of a family with two treated and two untreated ($0.25 \times 0.75 = 0.1875 < 0.25 = 0.5 \times 0.5$).

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

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

The numerator quantifies the “total amount of variation” identifying δ_{FE} by summing over the variation contributed by each group. This expression is similar to the weight given to each group in the FE estimate, except 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, $\text{Var}(D_{i,reference})$.

In our application, we report effective observations using two different reference groups. First, we normalize by the variation across units in a cross-section regression after controlling for reasonable g -level covariates, $\text{Var}(D_{i,reference})$. This gives the equivalent number of cross-sectional units that would produce the variation used in the FE regression. Second, we normalize by the within-family variation from units in two-unit groups. The difference between this and the number of individuals in switching families informs us about whether the presence of larger groups contributes additional identifying variation.²⁴

C. Consequences for Estimation: OLS Weights and FE Weights

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.

23. For example, Stuart et al. (2011) suggest that a 0.1–0.25 SD difference in propensity scores between the experimental and nonexperimental population may be too large to rely on extrapolation without further adjustments.

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

When treatment effects are heterogeneous, 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, and determines the magnitude of the effect of treatment. We allow for a different treatment effect for each value 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 $n_g \geq 2$, for example, siblings, is:

$$(4) \quad \hat{\delta}_{OLS} = \sum_{z \in \mathbb{Z}} \hat{\delta}_{z,OLS} \cdot \hat{\omega}_{z,OLS}$$

where $\hat{\omega}_{z,OLS}$ is the sample analogue to $\omega_{z,OLS}$:

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

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

The FE estimator for the same sample is:

$$(5) \quad \hat{\delta}_{FE} = \sum_{z \in \mathbb{Z}} \hat{\delta}_{z,FE} \cdot \hat{\omega}_{z,FE}$$

where $\hat{\omega}_{z,FE}$ is the sample analogue to $\omega_{z,FE}$:

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

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

Moving from OLS to FE, the $\hat{\delta}$ s change and also the $\hat{\omega}$ s change. The change in the $\hat{\delta}$ 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, $\hat{\omega}_z$, and how much is driven by the change in identification, $\hat{\delta}_z$, as:

$$(6) \quad \hat{\delta}_{FE} - \hat{\delta}_{OLS} = \sum_{z \in \mathbb{Z}} \underbrace{(\hat{\omega}_{z,FE} - \hat{\omega}_{z,OLS}) \cdot [\alpha \cdot \hat{\delta}_{z,FE} + (1 - \alpha) \cdot \hat{\delta}_{z,OLS}]}_{\text{Impact of } \Delta \text{ weighting}} + \sum_{z \in \mathbb{Z}} \underbrace{(\hat{\delta}_{z,FE} - \hat{\delta}_{z,OLS}) \cdot [\alpha \cdot \hat{\omega}_{z,OLS} + (1 - \alpha) \cdot \hat{\omega}_{z,FE}]}_{\text{OLS Bias}}$$

with $\alpha \in [0,1]$ being 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 $\hat{\omega}_z$ between OLS and FE, multiplied by an α -weighted average of the $\hat{\delta}_{z,OLS}$ and $\hat{\delta}_{z,FE}$.

Setting $\alpha=0$ in this term uses coefficients from a cross-sectional analysis 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 ω_z and δ_z , these two extremes can provide useful benchmarks to compare against the OLS and FE estimates, as we do in Section IV.C.²⁵ Equation 6 shows that SI matters more for the difference between OLS and FE when the gap in weights ($\hat{\omega}_{z,FE} - \hat{\omega}_{z,OLS}$) is correlated with treatment effects $\hat{\delta}_z$ (across elements of Z), and there is heterogeneity in $\hat{\delta}_z$.

A separate issue is that the FE estimate may not be unbiased for the ATE for switchers, $E[\hat{\delta}_{FE}] \neq \sum_z \delta_z \cdot Pr(z_g = z | S_g = 1)$. When there are heterogeneous treatment effects, unbiasedness would require the weights $\omega_{z,FE}$ to match with the expected population shares across values of Z . This will typically not be the case, as shown theoretically in Chernozhukov et al. (2013); Gibbons, Suárez, and Urbancic (2019); Imai and Kim (2019), and as we demonstrate in our empirical application below.

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 the proportion of the overall sample, the sibling sample, and the switcher sample comprised by families with one, two, three, four, or five or more kids. Moving from the sibling sample to the switchers sample, the share of the sample comprised of families with five or more kids roughly doubles (from 16.9 percent to 32 percent), while the share of the sample comprised of two-child families declines by 40 percent (from 34.5 percent to 21 percent.)

Panel B of Table 3 shows the variance in Head Start across the same family sizes and samples, netting out group fixed effects for the switcher sample. Going from the siblings sample to the switcher sample, the variance approximately doubles for all family sizes. This suggests that the change in the conditional variance across OLS and FE plays a relatively minor role in shifting weights across groups in our setting.²⁶

We then calculate $\hat{\omega}_{z,OLS}$ and $\hat{\omega}_{z,FE}$ in Panel C. Going from the sibling sample to the switchers sample, $\hat{\omega}_{2-child}$ declines by more than 25 percent, and $\hat{\omega}_{3-child}$ declines by 15 percent. Conversely, $\hat{\omega}_{5-child}$ nearly doubles from 0.134 to 0.243, and $\hat{\omega}_{4-child}$ families increases by more than 25 percent.

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

The treatment effect of Head Start also varies by family size in our applications. The first two columns of Panel A in Table 4 show 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

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

26. In Section VI, we provide additional evidence that “undoing” the conditional variance weighting makes little difference in this application.

Table 3*Change in Weighting of Regression Estimates across Sibling and Switcher Samples (PSID)*

	Number of Children in Family				
	1	2	3	4	5+
Panel A: Share of Sample					
All (no FFE)	0.123	0.273	0.238	0.147	0.134
Siblings sample (no FFE)	0.000	0.345	0.300	0.186	0.169
Switchers sample (FFE)	0.000	0.210	0.271	0.197	0.322
Panel B: Variance in Head Start					
All (no FFE)	0.089	0.104	0.121	0.127	0.132
Siblings sample (no FFE)	0.000	0.024	0.050	0.059	0.068
Switchers sample (FFE)	0.000	0.045	0.098	0.131	0.174
Panel C: Regression Weights					
All (no FFE)	0.171	0.257	0.284	0.117	0.101
Siblings sample (no FFE)	0.000	0.338	0.374	0.154	0.134
Switchers sample (FFE)	0.000	0.256	0.307	0.190	0.248

Source: Panel Study of Income Dynamics 1968–2011 waves.

Notes: This table shows the change in the composition of the PSID sample moving from all individuals and estimating a model without controls (“All (no FFE)”), to individuals that have at least one other sibling in the sample and estimating a model without controls (“Siblings Sample (no FFE)”), to individuals in families that have variation in Head Start attendance and estimating a model with family fixed effects (“Switchers sample (FFE)”). Panel A shows the share of individuals in each sample that come from a family with one child (zero siblings), two children, etc. Panel B shows the variance in Head Start for each family size and sample. For the switcher sample, 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_i and defined formally in Section III. The shares and regression weights do not sum to one for the “all sample” because this sample also includes an additional category of individuals who have an unknown number of siblings (due to a missing mother ID). These individuals account for 8.5 percent of the “all sample.”

families with five 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 because we find that the heterogeneity in family size survives the inclusion of other interactions, as we discuss in Section VI.

Table 4
Returns to Head Start by Family Size and Implications for Regression Estimates

	PSID			CNLSY			
	Some College			HS Grad		Idle	
	CX (1)	FE (2)	FE (3)	FE (3)	FE (4)	FE (5)	
Panel A: Effects by Family Size							
Head Start × 1 child family	0.169* (0.091)						
Head Start × 2 child family	0.038 (0.079)	-0.126 (0.099)	0.058 (0.050)		-0.075 (0.060)	-0.018 (0.025)	
Head Start × 3 child family	-0.030 (0.087)	0.152** (0.075)	0.042 (0.063)		-0.001 (0.071)	-0.073 (0.046)	
Head Start × 4 child family	-0.053 (0.100)	0.251*** (0.091)	0.135 (0.087)		-0.063 (0.118)	-0.042 (0.052)	
Head Start × 5+ child family	0.572*** (0.119)	0.348*** (0.126)	0.305*** (0.095)		-0.317** (0.132)	-0.161* (0.091)	
Head Start × Unknown child family	-0.099 (0.108)						
Observations	4,258	2,986	1,251		1,251	1,247	
Head Start switchers		213	668		668	668	
Effective obs. (individs. 2-person families)		235.9	644.3		644.3	644.3	
Effective obs. (CX individs.)		731.8	558.9		558.9	558.9	

(continued)

Table 4 (continued)

	PSID		CNLSY	
	Some College		HS Grad	Idle
	CX (1)	FE (2)	FE (3)	FE (4)
All	0.046			
Siblings	0.037	0.083	0.081	-0.071
Switchers	0.069	0.123	0.093	-0.077
				FE (5)

Panel B: Simulated Estimates across Samples Using Family-Size Regression Weights

Sources: Panel Study of Income Dynamics 1968–2011 waves and Children of the National Longitudinal Study of Youth.
 Notes: Panel A of this table shows the coefficients from regressions of outcomes 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 data source and specification varies across columns. Columns 1 and 2 use our main PSID sample, and the outcome is attainment of some college. Columns 3–5 use the CNLSY79 sample, and the outcomes are indicators for graduating from high school, being idle, and having a learning disability, respectively. Column 1 includes controls, but not mother fixed effects, and standard errors are clustered at the family ID level. Columns 2–5 include mother fixed effects, and standard errors are 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 Individuals)" is the equivalent number of cross-sectional units that provide the same amount of variation as switchers. "Effective Obs. (Individuals, 2-Person Families)" is the equivalent number of individuals in two-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_2 (defined in Section III), 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"). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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.²⁷ 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 two-person switching families. Hence, by including families with three or more children, on average, each observation is providing more variation than in a similar-sized sample of two-child families. Further, the variation is equivalent to 732 individuals in a cross-sectional regression. This is because there is relatively little variation in Head Start in the full sample.

The larger Head Start effects we document for big families are not specific to the PSID. Columns 3–5 of Table 4 show the effects of Head Start in the CNLSY by family size for high school graduation, idleness, and having a learning disability.²⁸ For each of these outcomes, the impact of Head Start for families with five or more children is at least twice as large as the impact for two- or three-child families. For high school graduation, we also see a large impact for four-child families, roughly double the impact for two- and three-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.

IV. Extrapolating from Identifying to Target Population

The difference between OLS and FE in the 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 . To give a few examples, the target population could be a nationally representative sample, which is a common group of interest in applied work, or, more relevant for our application, it could include families that are eligible for a means-tested program or families that have at least one participant in a means-tested program.

A. Assumptions and Proposition

The reweighting method relies 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.

Second, we assume that conditional on observables, the true treatment effect is independent of a group’s switching or target status. We refer to this as “conditional fixed effect

27. 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 $[V(D_i|g) \cdot (n_g - 1)]/n_g = [0.5^2 \cdot (2 - 1)]/2 = 0.125$.

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

ignorability” (CFEI). We use two propensity scores constructed from a vector of observable group characteristics, \mathbf{X}_g , as the conditioning variables: $P_x := Pr[S_g = 1 | \mathbf{X}_g = \mathbf{x}]$ is the population propensity to be a switching group, and $Q_x := Pr[T_g = 1 | \mathbf{X}_g = \mathbf{x}]$ is the population propensity to be in the (researcher-determined) target group.²⁹

Assumption 2: Conditional Fixed Effect Ignorability (CFEI)

$$(7) \quad E[Y_i(1) - Y_i(0) | S_g, P_x, Q_x] = E[Y_i(1) - Y_i(0) | P_x, Q_x]$$

$$(8) \quad E[Y_i(1) - Y_i(0) | T_g, P_x, Q_x] = E[Y_i(1) - Y_i(0) | P_x, Q_x]$$

Assumption 2 is a strong assumption. CFEI eliminates, for example, a second type of Roy (1951)–type selection, whereby switchers have an unobserved quality that increases the effectiveness of treatment compared to observationally equivalent non-switchers. This assumption is required in order to extrapolate treatment effects from switchers to nonswitchers. Analogous assumptions are made in the IV literature extrapolating from compliers to a target population (Angrist and Fernandez-Val 2013; Aronow and Carnegie 2013). The plausibility of Assumption 2 will depend on the context of the specific application.

In our application, this assumption may be reasonable because observables strongly predict both switching and target status. We have shown that 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, are strong predictors of 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.

However, there may be unobservables that determine switching or target status, which are also correlated with treatment effects. To address this concern, we develop two testable implications of Assumption 2 at the end of this subsection.

In contexts where this assumption may be too strong, the reweighting procedure can still help to reduce bias, since balancing on observables will likely reduce the degree of bias that is driven by mismatch on unobservables. Andrews and Oster (2019) develop this idea formally in the context of extrapolating from an experiment to a larger target population. This improvement relies on the assumption that the direction of bias from selection on observables is the same as the direction of bias from selection on unobservables, which may be plausible in many cases.³⁰

29. For instance, if the target group is “all individuals,” $Q_x = 1$ for each individual, while if the target group is “Head Start participants,” Q_x is the population probability that an individual in a group with $\mathbf{X}_g = \mathbf{x}$ is a Head Start participant. Note that the target group is “all individuals.” CFEI simplifies to the assumption that treatment effects are independent of switching status, conditional on the probability of being a switcher, P_x .

30. For example, we show that larger families are likely to have more variation and larger treatment effects. Such families may also be more motivated (or better equipped) to seek out information about public programs, which would also predict greater variation and larger treatment effects.

Assumption 3: Correct Propensity Score Specification

$$(9) \ Pr(S_g = 1 | \mathbf{X}_g) = F(\theta; \mathbf{X}_g)$$

$$(10) \ Pr(T_g = 1 | \mathbf{X}_g) = G(\chi; \mathbf{X}_g)$$

Third, we assume that the propensity scores that we estimate have the correct functional form, with $F(\cdot)$ and $G(\cdot)$ known, and θ and χ parameters to be estimated. In our application, we model $F(\cdot)$ and $G(\cdot)$ jointly as a multinomial logit.

Assumption 4: Overlap

$$(11) \ \text{If } Q_x > 0, \text{ then } P_x > 0$$

Fourth, we require a positive probability of being a switcher for each value of Q_x 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 one-unit groups (“singletons”) in the switcher sample, but they are present in some of our application target populations,³¹ we cannot allow treatment effects for singletons to be outside of the support of the treatment effects of switchers. This 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 two or fewer units, which allows us to extrapolate treatment effects for singletons from two-unit groups.³² 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 reweighted FE estimator for target population t

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

with $\widehat{w}_{g(i)}^t$ our estimate of $w_{g(i)}^t$,

$$(13) \ w_{g(i)}^t := \frac{Q_x \cdot Pr[S = 1]}{P_x \cdot Pr[T = 1]}$$

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

31. Singletons comprise 6 percent of Head Start participants in the CNLSY, and 18 percent of Head Start participants in the PSID.

32. We implement this by including an indicator for “one- or two-child families” in \mathbf{X}_g together with indicators for other family sizes. Alternatively, the target group can be defined to include only families that are ever switchers, such as “siblings” or “multi-child Head Start families.”

The proof is in [Online 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 other terms in the weight, $Pr[S=1]$ and $Pr[T=1]$, are constant across groups. 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.³³ Because the weights increase as \hat{P}_x gets close to zero, for implementation we suggest directly examining the estimated weights and using caution if there are values of \hat{P}_x that are extremely small.

Testable Implications

Conditional fixed effect ignorability requires that treatment effects should be balanced across T_g , conditional on P_x and Q_x (Equation 8). Hence, switchers in the target population should on average have the same treatment effects as switchers in the nontarget population once observables are balanced across the two groups. This is potentially testable if some switchers are not in the target population (that is, if the target is not “everyone,” “multi-unit groups,” or otherwise contains the set of switchers G_S). 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 could test whether treatment effects are different for participants and nonparticipants, rural and urban individuals, etc.

Conditional fixed effect ignorability also requires that treatment effects should not vary across switchers and nonswitchers (Equation 7). This could be violated, for example, if families with a lower return to Head Start make less of an effort to have any of their children attend or if families with a higher return to Head Start make a greater effort to have all of their children attend. Because we never observe the treatment effects of nonswitchers, we can not directly test for this type of violation. However, this type of violation generates a testable implication—if families systematically choose the number of children to attend Head Start based on private information, we should also expect to see this pattern *within* switchers. In particular, if switchers act on private information about the size of treatment effects, then switcher families that have a higher fraction of children that attend Head Start should experience systematically larger gains. Thus, a useful diagnostic test is whether treatment effects vary with the number or fraction of treated units, holding constant the size of the family.³⁴

We implement both of these tests for our setting in Section VI.B.1.

B. Reweighting Methodology

In order to implement this reweighting strategy, we first need to obtain estimates of P_x and Q_x . These two elements can be calculated using a multinomial logit model where the outcomes are the four possible combinations of having $S_g = 1$ or $S_g = 0$, and $T_g = 1$

33. See [Online Appendix A](#) for a simple derivation of the weights.

34. Controlling for group size allows us to distinguish whether families have a higher fraction attend due to *selection* or due to other factors, such as the mechanical relationship between family size and probability of being a switching family, or unmeasured variables that are correlated with family size.

or $T_g=0$. \hat{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 group. \hat{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 group.

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. Aggregating $\widehat{w}_{g(i)}^t$ to the group level and performing a normalization,

we obtain the two-step estimation weights, $\hat{s}_g^t = \frac{\widehat{w}_{g(i)}^t \cdot n_g}{\sum_{g' \in G_s} \widehat{w}_{g(i)'}^t \cdot n_{g'}} = \frac{\hat{Q}_x \cdot n_g}{\sum_{g' \in G_s} \frac{\hat{Q}_x}{P_x} \cdot n_{g'}}$. The two-step ATE is then:

$$(14) \quad \widehat{\delta}_{2step}^t = \sum_{g \in G_s} \hat{s}_g^t \cdot \hat{\delta}_g$$

A second approach is to obtain the ATE in a single step using “in-regression weights.” For this, we need to incorporate an additional element to our p-score weights that “undoes” the conditional variance weighting of FE: $v_g = \text{Var}(D_i | g(i)=g)^{-1}$ (Gibbons, Suárez, and Urbancic 2019).³⁵ The ATE can then be estimated from a one-step regression using $\widehat{w}_{g(i)}^t \cdot v_g$ as regression weights.³⁶

Our results in Section IV.A assume simple random sampling. In our application, we adjust our reweighting procedure to incorporate survey weights. First, we weight our multinomial logit using survey weights. The resulting \widehat{w}^t then has the interpretation of translating between “population switchers” and “population target observations.” Second, we multiply $\widehat{w}_{g(i)}^t$ by survey weights for both the one- and two-step reweighting procedures. We obtain standard errors for all estimates by bootstrapping.

C. 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 Equation 6), we obtain a weighted coefficient of 0.069, shown in the bottom row of Table 4. This implies that approximately one-third of the change from OLS to FE [(0.069 – 0.049)/(0.12 – 0.049)] is driven by the change in weighting across family sizes, with the other two-thirds driven by change in identifying variation. Further, reweighting the FE estimates using the OLS weights ($\alpha=1$ in Equation 6) produces a coefficient is 0.083. This implies that the imbalance in family size alone causes the FFE estimate to be 50 percent higher than the estimates without FE.

35. This variance is computed using the population formula, (dividing by n_g), rather than the sample formula (dividing by n_g-1).

36. In the special case where the target group is “switchers,” then $P_x = Q_x$, $\widehat{w}_{g(i)}^t = 1$, and the one-step weight simplifies to v_g , so that we recover the ATE for switchers.

D. Monte Carlo Experiments

We perform a Monte Carlo analysis to examine the properties of our proposed reweighting estimators. We use naturally occurring 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 (assumed to be known to the researcher), which determines the covariates that we use to generate the propensity score.

We generate the data for the Monte Carlo as follows. We construct an untreated outcome for each individual by running a linear probability model of attainment of “some college or more” on demographic variables, income during childhood, and parental education. We then use the estimates from this model to generate a prediction of a continuous probability that an individual completes some college, X_{ig} .³⁷ 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 and varies across different models of heterogeneity. We scale Y_{ig}^* to ensure that these probabilities lie within the range [0,1]. We then randomly generate the binary outcome variable as $Pr(Y_{ig} = 1) = Y_{ig}^*$.

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 four or more siblings), and $\beta_{ig} = 0$ for small families (three 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: $\beta_{ig} = 0.08 \left[1 - \frac{X_{ig} - \bar{X}_{ig}}{SD(X_{ig})} \right] \frac{1}{3}$, with \bar{X}_{ig} and $SD(X_{ig})$ the sample mean and standard deviation of X_{ig} . This produces a treatment effect that is larger for individuals with a lower baseline probability 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 5th, 20th, 50th, 80th, and 95th percentiles of X_{ig} . The latter model presumes that the researcher has some intuition that the treatment effect or selection into identification may vary nonlinearly 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.³⁸ The FE estimate is the same for all target populations. We consider four target populations: (i) individuals in Head Start switching families,³⁹ (ii) all siblings, (iii) all individuals in the sample (including singletons), and (iv) all Head Start participants. We multiply all estimates by 1,000 for easier readability.

Panel A of Table 5 presents results for the model with constant treatment effects. In this setting, the average treatment effect is the same for all target populations, all estimators are unbiased, and the FE model is the minimum variance estimator. The reweighting estimators have mean squared errors 3–20 percent larger than for OLS.

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

38. Post-regression and in-regression reweighting produce the same results.

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

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

	True ATE	Bias:		Ratio: MSE of Reweight to MSE of FE
		FE	Reweight	
Panel 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
Panel 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
Panel 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
Panel 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 IV.D. Panel A shows results where Head Start has a constant treatment effect (TE) for all individuals. Panel B shows results where Head Start (HS) has no effect on individuals from small families (three or fewer children) and a large effect for families with many children (four or more children). Panels C and D show results where treatment effects 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. * $p < 0.01$.

Panel B of Table 5 presents results for the model with zero treatment effect for small families, and large treatment effects for large families (four or more children). 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.⁴⁰

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 percentage points and −0.08 percentage points on a base of nine percentage points), but has much larger bias for the remaining targets. Panel C shows that the regression reweighting estimator that 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.⁴¹

Panel D shows that when we reestimate 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, it performs 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.

V. Extensions

A. 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-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) that could also enable the recovery of the 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.⁴² The weighted

40. In results not reported, we 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 percentage points, relative to the two to three percentage point bias in FE) for the “siblings” and “all” target groups.

41. 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.

42. See [Online Appendix A.1.1](#) for a formalization of this assumption and an extension of Proposition 1 using extrapolation.

average of estimated treatment effects for $P_x > 0$ and extrapolated effects for $P_x = 0$ would then give the ATE for the target group.

B. Unit i Covariates

Thus far, we have ignored covariates in our models. However, researchers may want to include individual-level covariates C_i (that vary across i units within a group) in their models in order to make Assumption 1 more reasonable, improve precision of estimates, and 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 .⁴³ Thus, treatment effects can also be estimated for residual switchers; however, identifying variation comes from within-family variation in C_i , not D_i . In [Online Appendix A.3](#) we discuss additional considerations related to residual switchers, including how one can quantify the contribution of residual switchers to estimate, and how our key assumptions and proposition can be extended to accommodate C_i .

C. Reweighting Misspecified FE Models

We have also focused on cases where Assumption 1 (CIA) holds, and the within-group comparisons form unbiased estimates of group-level treatment effects. If instead there are violations of Assumption 1, then reweighting estimates can affect the resulting bias. The impact of reweighting on the misspecification bias will depend on the covariance (across groups) of group-specific bias, $bias_g$, and the impact of the reweighting procedure on the weight given to each group: $E_g \left[\left(w_g^t - \omega_{g,FE} \right) \cdot bias_g \right]$. For example, suppose that smaller-sized groups had a more positive bias and that the reweighting increased the weight of these groups. This would make the misspecification bias more positive and in principle could possibly lead to a net increase in the overall bias. This suggests that researchers may want to consider violations of CIA by group characteristics (for example, group size) before reweighting.

D. Nonlinear Functional Form

Next, we relax the linear functional form assumption. One reason this may make a difference is that conditional or fixed effect logit models use only “double switchers,” families with variation in *both* the outcome variable and the treatment variable. In [Online 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.

E. 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

43. See [Online Appendix A.3](#) for a formalization of this.

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.

VI. Effects of Head Start

A. Data and Replication of Garces, Thomas, and Currie (2002) 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 Garces, Thomas, and Currie (2002) and Deming (2009).

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 Garces, Thomas, and Currie (2002). 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 Garces, Thomas, and Currie (2002) in [Online 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 Garces, Thomas, and Currie (2002). We find large (23 percentage points) and significant effects of Head Start on the probability that whites attain some college and large point estimates (9.3 percentage points) 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.⁴⁴

For the remaining analyses from here, we use a sample that substantially expands and modifies the Garces, Thomas, and Currie (2002) 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 Garces, Thomas, and Currie (2002) 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

44. In some subsamples, we 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 undiagnosable) differences in specification and/or dataset construction.

Head Start affects children's lives. We follow the established practice of distilling the measures to summary indexes to lessen problems with multiple hypothesis testing (see, for example, Anderson 2008; Kling, Liebman, and Katz 2007; Hoynes, Schanzenbach, and Almond 2016). We create four indexes 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: nonsmoking, report of good health, and negative of mean BMI.⁴⁵

The process of creating each index follows the procedure described by Kling, Liebman and Katz (2007). In particular, we standardize each component of the index by subtracting the mean outcome for nontreated children, defined as children who did not attend any form of preschool, and then dividing the result by the standard deviation of the outcome for nontreated children. The summary index takes a mean of these standardized measures.⁴⁶

[Online Appendix Table B.3](#) 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/nonparticipants, and for the sample of individuals with siblings. We present the means of the analyzed outcomes in [Online Appendix Table B.4](#).⁴⁷

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 who took part in the NLSY79 survey. The sample we use includes all children who were at least 4 years old by 1990.

B. 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 Garces, Thomas, and Currie (2002) and Deming (2009), we estimate:

$$(15) \quad Y_{ig} = \alpha + \beta_1 \text{HeadStart}_{ig} + \beta_2 \text{OtherPreSchool}_{ig} + \mathbf{X}_{ig} \gamma + \delta_g + \varepsilon_{ig}$$

45. See [Online Appendix Table B.5](#) for descriptive statistics of the inputs to the indexes.

46. 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 nonpreschool participants.

47. [Online Appendix Table B.5](#) includes summary statistics for the inputs to the summary indexes. [Online Appendix Tables B.6, B.7, and B.8](#) contain the number of observations for each outcome and control variable in the analysis.

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.⁴⁸ δ_g is a mother fixed effect that 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 that 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.⁴⁹

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.⁵⁰ 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: individuals eligible for Head Start based on family income between ages two and five;⁵¹ all Head Start participants, and all siblings. 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 three, and income at age four, and indicators for family size (grouping together one- and two-child families).⁵² We include results for the post-regression weighting method; results are qualitatively similar when we use in-regression weighting.

C. Evidence on Model Assumptions: Identifying and CFEI

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

48. Since Head Start only became available in 1965, we recode Head Start attendance to be “other preschool” for the 1961 and older cohorts.

49. 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 four; four-knot splines in annual family income for each age zero, one, and two; a fourth spline based on average family income between ages three and six; indicators for mother’s employment status at ages zero, one, and two; and household size at age four. For the CNLSY, these include: health conditions before age five; PPVT test score at age three; measures of birth weight; measures of mother’s health and health behaviors; mother’s working behavior and income prior to age four; indicator for being firstborn; participation in Medicaid; relative care; and indicators for early care types.

50. 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.

51. An individual is considered Head Start eligible if at any point between the ages of two and five their family income was below 150 percent of the poverty level in order to account for our imperfect ability to observe reportable income.

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

suggests that the magnitude of selection may be small. In [Online Appendix Table B.9](#), 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 (and omitting these characteristics as controls where relevant). For the white sample that forms our focus, there are few statistically significant correlations, which suggests that the assumption may be reasonable.⁵³

As a first test of CFEI (Assumption 2), [Online Appendix Table B.10](#) analyzes whether treatment effects vary across observationally comparable families in the target and nontarget populations. Recall that CFEI implies that conditional on P_x and Q_x , treatment effects should be independent of observable characteristics, including membership in the target population. To implement this test, 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 (and therefore treatment effects, if CFEI holds) between target and nontarget switchers.⁵⁴ When target status varies within family, another way of viewing this test is as asking: Is the treatment effect related to the share of individuals in the family that are in the target? This test passes with no sign of systematic differences across target and nontarget individuals across all outcomes, two data sets, and two different target populations (Head Start eligible and Head Start participants), although the standard errors are often large.⁵⁵

As a second, related, test, we examine whether the impact of Head Start varies with the fraction of children that attend Head Start in a family (which could signal that switchers have private information about the gains from Head Start). [Online Appendix Tables B.11 and B.12](#) show the results for the PSID and CNLSY samples, respectively. In the PSID, we find no significant interaction with the fraction of children that attend Head Start. In the CNLSY, we find that five-child families that have more children attend Head Start have smaller improvements in high school and larger improvements in learning disabilities, although this relationship appears to be driven by a very small number of individuals. Averaging across all family sizes, however, we find no significant relationship between the fraction that attend Head Start and treatment effects. This lends further support to our assumption that it is appropriate to extrapolate from switchers to nonswitchers.⁵⁶

Finally, extrapolation from switchers to the target population also requires that switchers have appropriate representation over the distribution of P_x and Q_x .

53. For the Black sample, participation in Head Start is correlated with a greater likelihood of having higher income at age one and lower income at age two, 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.

54. For target individuals the weights are $1/\hat{Pr}[T_i = 1, S_g = 1 | X_{ig}]$, and for nontarget individuals the weights are $1/\hat{Pr}[T_i = 0, S_g = 1 | X_{ig}]$.

55. We have run analogous models at the family level, which give qualitatively similar results.

56. If we take the negative slope on the fraction attending at face value (and disregard the noise in the estimates), it suggests that for the target population of HS participants, our treatment effects might be too large. In contrast, for a target population that has many nonswitcher nonparticipants, our estimated treatment effect might be too small. If there are systematic differences, one approach to correct for this would be to include fraction participating as an additional covariate, with dummies that span bins that cover both switching and nonswitching families, although, in that case, it may also be prudent to place less emphasis on the reweighted estimates.

Assumption 4 (overlap) implies that this will hold asymptotically, but for a finite sample this should be directly checked. Thus, we compute the fraction of target observations that lie within the convex hull of P_x and Q_x for the switcher observations.⁵⁷ [Online Appendix Figure B.4](#) shows this graphically for the “Head Start participant” target. In total, 95 percent of Head Start participants lie within the convex hull of the switcher sample. Of those that are outside, most are close to the convex hull. We interpret this magnitude of violation of the overlap assumption as mild enough to disregard in our subsequent analysis.⁵⁸

D. Head Start Results

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 Garces, Thomas, and Currie (2002), Column 2 presents the results using our expanded sample, and Columns 3–5 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 percentage points (SE: 0.053) using the baseline FFE model. This estimate is 57 percent smaller than the estimate reported in Garces, Thomas, and Currie (2002) of 0.281 (SE: 0.108).⁵⁹ The standard errors are also roughly 50 percent smaller, corresponding to the roughly tripling of sample size (2,986 compared with 1,036).

As we foreshadowed, 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 two- to three-child families (Panel A) and four-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. This figure also shows that there are no weights that are unduly large, so we do not trim any observations based on \hat{P}_x .

The reweighted estimates of the impact of Head Start for the eligible, participant, and sibling populations are 0.052, 0.021, and 0.064, respectively, and are all statistically insignificant.⁶⁰ Setting aside the lack of precision in the estimates, these represent

57. Traditional methods based on a single propensity score involve checking the overlap of the propensity score distributions. We extend this idea to two dimensions.

58. [Online Appendix Figure B.5](#) shows an analogous graph for the Head Start eligible sample. Here, 90 percent of target individuals are within the convex hull, and all of those outside are very close to it.

59. We show in the [Online Appendix](#) that this discrepancy is not due to faulty replication of the Garces, Thomas, and Currie (2002) estimates in a smaller sample. We estimate a coefficient of 0.232 (SE: 0.094) for this sample and outcome in our replication.

60. This reflects both the reweighting of heterogeneous effects across family sizes that we documented earlier, as well as slight changes to the family-size specific treatment effects (due to changes in the weighting of families that are less-represented along other covariates within a family size).

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

	FFE		Reweighted ATE, Target =				Diff. b/w FFE and Participant ATE
	Garces, Thomas, and Currie/Deming	Expand Sample/Replicate	HS Eligible	Participants	Siblings		
Panel A: Some College (PSID)							
Head Start	0.281** (0.108)	0.120** (0.053)	0.052 (0.064)	0.021 (0.059)	0.064 (0.061)	0.099** (0.032)	
Y mean in target		0.556	0.387	0.437	0.556		
Panel B: Economic Sufficiency Index, Age 30 (PSID)							
Head Start		-0.023 (0.102)	-0.071 (0.101)	-0.040 (0.099)	0.021 (0.113)	0.017 (0.066)	
Y mean in target		0.213	-0.198	-0.485	0.213		
Panel C: High School Graduation (CNLSY)							
Head Start	0.086*** (0.031)	0.085*** (0.030)	0.033 (0.042)	0.048 (0.037)	0.020 (0.044)	0.037* (0.023)	
Y mean in target		0.776	0.734	0.766	0.776		
Panel D: Idle (CNLSY)							
Head Start	-0.071* (0.038)	-0.072* (0.037)	-0.061 (0.045)	-0.055 (0.042)	-0.067 (0.050)	-0.017 (0.026)	
Y mean in target		0.197	0.221	0.201	0.197		

(continued)

Table 6 (continued)

	FFE		Reweighted ATE, Target =			Diff. b/w FFE and Participant ATE
	Garces, Thomas, and Currie/Deming	Expand Sample/Replicate	HS Eligible	Participants	Siblings	
Panel E: Learning Disability (CNLSY)						
Head Start	-0.059*** (0.020)	-0.059*** (0.021)	-0.031 (0.026)	-0.042* (0.022)	-0.040 (0.026)	0.017 (0.015)
Y mean in target		0.051	0.055	0.041	0.051	
Panel F: Poor Health (CNLSY)						
Head Start	-0.070*** (0.026)	-0.069*** (0.026)	-0.063* (0.037)	-0.067** (0.034)	-0.050 (0.038)	-0.003 (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 Garces, Thomas, and Currie (2002) 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 3–5 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). Sample size is $N=2,986$ for the expanded sample, and 1,036 for Garces, Thomas, and Currie (2002). Standard errors obtained by bootstrapping. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

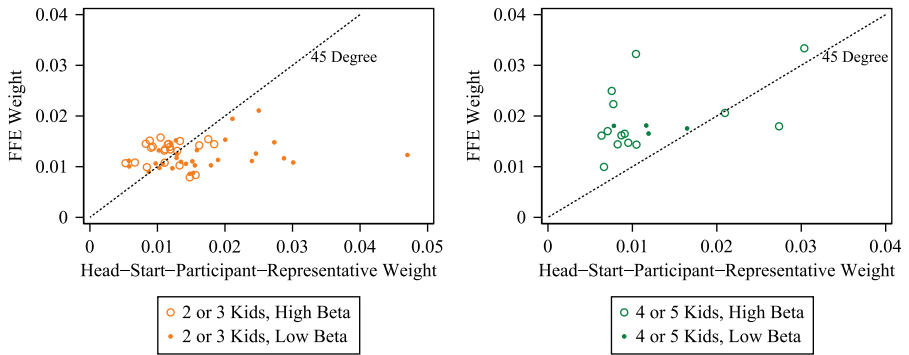


Figure 3

Family Fixed Effects Weights and Head Start Participant Representative Weights by Family Size and Some College β (PSID White Sample)

Source: Panel Study of Income Dynamics 1968–2011 waves.

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 two to three children or four 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.

moderately large impacts relative to the 43.7 percent average rate of college going among Head Start eligible children. But relative to the FFE coefficient, these effects imply a 46–83 percent smaller impact on college attendance. Putting these estimates in broader perspective, they are 45–91 percent smaller than the unadjusted estimates for *all* participants from other FFE studies (Bauer and Schanzenbach 2016; Deming 2009) and 51 percent smaller than the estimate from the county rollout of Head Start (Bailey, Sun, and Timpe 2021), although the lower end of the confidence intervals for these estimates includes our ATE.

Panel B of Table 6 presents results for the economic sufficiency index in the PSID. Our FFE estimate shows a statistically insignificant 0.023 SD decline in this index associated with Head Start. When we reweight the effects, we find slightly larger negative effects for Head Start eligible children and Head Start participants and a positive effect (0.02 SD) for siblings. It bears emphasizing, though, that the results are not precisely estimated, such that the 95 percent 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 allows for a Head Start induced improvement of 0.15 SD or a reduction of 0.23 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 percentage point increase in high school graduation ($p < 0.01$), a 7.2 percentage point decline in idleness ($p < 0.01$), a 5.9 percentage point decline in having a learning disability ($p < 0.01$), and a 6.9 percentage point decline in reporting poor health ($p < 0.01$). The reweighted estimate for participants for high school is 44 percent smaller and not statistically significant. We also see substantial 24 percent and 28 percent 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 percent 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 percent and 10 percent levels, respectively. The remainder of the outcomes are more imprecisely estimated, so we cannot reject that the reweighted estimate is the same as the FFE estimate.

Returning to the PSID, [Online Appendix Tables B.13 and B.14](#) 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 for these outcomes. 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 is intuitive 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 -0.27 SD, but reweighting for the Head Start participant population changes this estimate to -0.34 . In contrast, the coefficients are relatively stable for Blacks.⁶¹

We explore other reweighting strategies in [Online Appendix Tables B.15 and B.16](#). Reweighting using linear extrapolation of treatment effects to singletons in [Table B.15](#) produces qualitatively similar results to the baseline reweighting.⁶² [Table B.16](#) 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.

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

61. 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 ($SE = 0.073$).

62. We have also explored excluding singletons altogether from the target. The estimates for nonsingleton Head Start participants and nonsingleton Head Start eligible children typically lie between the reweighted estimates for siblings and Head Start participants.

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, so our CNLSY results are to some degree an out-of-sample confirmation of this pattern.

We examined whether the larger coefficients for larger family sizes in Table 4 are driven by family size standing in for other covariates. In [Online Appendix Table B.17](#), 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 heterogeneity along other covariates. We 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.

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, and heterogeneity by race, gender, and cohort in [Online Appendix C](#). We do not find any systematic evidence of effects on any of these outcomes, or important heterogeneity along these dimensions.

VII. 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 estimate. Since returns to migration may also be heterogeneous by family characteristics, it may be useful to reweight the estimates from Collins and Wanamaker (2014) as well 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 eight-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 $Pr(S_g = 1)$, nonswitcher schools probably also have a different set of school resources (for example, share of highly experienced teachers) and student composition (for example, 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 nonswitchers or high-probability-DV nonswitchers.

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 that are less negative than those estimated with other strategies. The authors report that 91 percent of the county pairs in the data have variation in the minimum wage at some point during the analysis, but states with more border counties and that 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 percent. 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 (for example, district FE). For example, Shah and Steinberg (2017) employ this strategy and find that a positive rain shock (top 20 percent 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 four years of analysis. These nonswitching districts may be located in a different geography or have distinct industrial composition or population characteristics, which could in turn affect the elasticity of school attendance. Hence, extrapolating from switcher to nonswitcher districts may require reweighting strategies such as those we propose.

Some recent work notes empirical issues related to SI. In the context of estimating returns to experience for teachers, Wiswall (2013) observes that teacher-FE models, paired with modest panel lengths and experience categories such as “4+ years,” imply that the coefficients on these categories will be solely identified from relatively new teachers and lead to misestimation of the returns to experience. Spears, Coffey, and Behrman (2019) raise SI concerns for FFE models used by Jayachandran and Pande (2017) to examine the relationship between birth order and child height in the Demographic and Health Surveys. Since these outcomes are only observed for those under five, Spears, Coffey, and Behrman (2019) argue that a FE design selects for families with shorter birth spacing, which in turn may have more negative impacts of birth order.

As researchers using FE consider the relevance of SI for a particular setting, we also note that greater degrees of SI may, on the margin, favor a design that uses a more coarse level of fixed effects in order to increase the number of switchers. For example, Geruso and Spears (2018) examine the impacts of heat on infant mortality. They use district-by-calendar-month FEs in one specification, and village-by-calendar-month FEs as an alternative. The latter specification, with smaller groups, may have greater SI concerns. It also has larger standard errors, and the authors suggest that the coarser aggregation is their preferred model. The reweighting methods we propose could be employed to mitigate the SI concerns in applications that use more narrow FEs.

Overall, 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.

VIII. 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 (nonswitchers). 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 ATEs 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 2.1 percentage point increase in the likelihood of attending some college for Head Start participants and a 5.2 percentage point increase for those who are Head Start eligible. The ATE estimate for participants is 83 percent smaller than the FFE estimate, a difference that is statistically significant at the 5 percent 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. In the CNLSY, for high school graduation we find that the reweighted estimate for participants (4.8 percentage points) is 44 percent smaller than the FFE estimate, a difference that is statistically different at the 10 percent 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 of 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 (for example, 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." NBER Working Paper 18429. Cambridge, MA: NBER.
- Almond, Douglas, Kenneth Y. Chay, and David S. Lee. 2005. "The Costs of Low Birth Weight." *Quarterly Journal of Economics* 120(3):1031–83.
- 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(484):1481–95.
- 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." NBER Working Paper 22721. Cambridge, MA: NBER.
- Andrews, Isaiah, and Emily Oster. 2019. "A Simple Approximation for Evaluating External Validity Bias." *Economics Letters* 178:58–62.
- Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66(2):249–88.
- Angrist, Joshua D., and Ivan Fernandez-Val. 2013. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." In *Advances in Economics and Econometrics: Tenth World Congress*, Econometric Society Monograph 3, ed. Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 401–34. Cambridge, UK: Cambridge University Press.

- Angrist, Joshua, and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton 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. 2021. "Prep School for Poor Kids: The Long-Run Impacts of Head Start on Human Capital and Economic Self-Sufficiency." *American Economic Review* 111(12):3963–4001.
- Barr, Andrew, and Chloe R. Gibbs. 2022. "Breaking the Cycle? Intergenerational Effects of an Antipoverty Program in Early Childhood." *Journal of Political Economy* 130:3253–85.
- Bates, Michael David, Katherine E. Castellano, Sophia Rabe-Hesketh, and Anders Skrondal. 2014. "Handling Correlations between Covariates and Random Slopes in Multilevel Models." *Journal of Educational and Behavioral Statistics* 39(6):524–49.
- Bauer, Lauren, and Diane Whitmore Schanzenbach. 2016. "The Long-Term Impact of the Head Start Program." Washington, DC: The Hamilton Project, Brookings Institute.
- Bayer, Patrick, Randi Hjalmarsson, and David Pozen. 2009. "Building Criminal Capital behind Bars: Peer Effects in Juvenile Corrections." *Quarterly Journal of Economics* 124(1):105–47.
- 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." *Quarterly Journal of Economics* 122(1):409–39.
- Borusyak, Kirill, and Xavier Jaravel. 2017. "Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume." Unpublished.
- 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):169–82.
- 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." Unpublished.
- Cameiro, 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):135–73.
- 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):211–28.
- Carrell, Scott E., Mark Hoekstra, and Elira Kuka. 2018. "The Long-Run Effects of Disruptive Peers." *American Economic Review* 108(11):3377–415.
- Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The Effect of Minimum Wages on Low-Wage Jobs." *Quarterly Journal of Economics* 134(3):1405–54.
- Chaisemartin, Clement, and Xavier D'Haultfoeille. 2019. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." Unpublished.
- Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey. 2013. "Average and Quantile Effects in Nonseparable Panel Models." *Econometrica* 81(2):535–80.
- Chetty, Raj, and Nathaniel Hendren. 2018a. "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects." *Quarterly Journal of Economics* 133(3):1107–62.
- . 2018b. "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates." *Quarterly Journal of Economics* 133(3):1163–228.
- Chorniy, Anna V., Janet Currie, and Lyudmyla Sonchak. 2018. "Does Prenatal WIC Participation Improve Child Outcomes?" NBER Working Paper 24691. Cambridge, MA: NBER.
- 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):220–52.
- 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):1691–701.

- Currie, Janet, and Maya Rossin-Slater. 2013. "Weathering the Storm: Hurricanes and Birth Outcomes." *Journal of Health Economics* 32(3):487–503.
- Currie, Janet, and Duncan Thomas. 1995. "Does Head Start Make a Difference?" *American Economic Review* 85(3):341–64.
- De Haan, Monique, and Edwin Leuven. 2020. "Head Start and the Distribution of Long-Term Education and Labor Market Outcomes." *Journal of Labor Economics* 38(3):727–65.
- Deming, David. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1(3):111–34.
- Dube, Arindrajit, T. William Lester, and Michael Reich. 2010. "Minimum Wage Effects across State Borders: Estimates Using Contiguous Counties." *Review of Economics and Statistics* 92(4):945–64.
- . 2015. "Minimum Wage Shocks, Employment Flows, and Labor Market Frictions." *Journal of Labor Economics* 34(3):663–704.
- 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):3921–55.
- Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2016. "Sources of Geographic Variation in Health Care: Evidence From Patient Migration." *Quarterly Journal of Economics* 131(4):1681–726.
- Garces, Eliana, Duncan Thomas, and Janet Currie. 2002. "Longer-Term Effects of Head Start." *American Economic Review* 92(4):999–1012.
- Geruso, Michael, and Dean Spears. 2018. "Heat, Humidity, and Infant Mortality in the Developing World." NBER Working Paper 24870. Cambridge, NBER.
- Gibbons, Charles E., Serrato Juan Carlos Suárez, and Michael B. Urbancic. 2019. "Broken or Fixed Effects?" *Journal of Econometric Methods* 8(1):20170002.
- Gibbs, Chloe, Jens Ludwig, and Douglas L. Miller. 2013. "Head Start Origins and Impacts." In *Legacies of the War on Poverty*, ed. Martha J. Bailey and Sheldon Danziger, 51–69. New York: Russell Sage Foundation.
- Goodman-Bacon, Andrew. 2018. "Difference-in-Differences with Variation in Treatment Timing." NBER Working Paper 25018. Cambridge, MA: NBER.
- Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics* 115(4):1239–85.
- 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.
- Imai, Kosuke, and In Song Kim. 2019. "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 63(2):467–90.
- Jayachandran, Seema, and Rohini Pande. 2017. "Why Are Indian Children so Short? The Role of Birth Order and Son Preference." *American Economic Review* 107(9):2600–2629.
- Johnson, Rucker C., and C. Kirabo Jackson. 2019. "Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public School Spending." *American Economic Journal: Economic Policy* 11(4):310–49.
- 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):261–91.
- Lochner, Lance, and Enrico Moretti. 2015. "Estimating and Testing Models with Many Treatment Levels and Limited Instruments." *Review of Economics and Statistics* 97(2):387–97.
- Loken, Katrina 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." *Quarterly Journal of Economics* 122 (1):159–208.
- Pages, Remy J.-C., Dylan J. Lukes, Drew H. Bailey, and Greg J. Duncan. 2020. "Elusive Longer-Run Impacts of Head Start: Replications within and across Cohorts." *Educational Evaluation and Policy Analysis* 42(4):471–92.
- 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):135–46.
- 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):527–61.
- Sloczynski, Tymon. 2018. "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands." Unpublished.
- Spears, Dean, Diane Coffey, and Jere Behrman. 2019. "Birth Order, Fertility, and Child Height in India and Africa." NBER Working Paper 12289. Cambridge, MA: NBER.
- 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):369–86.
- Suri, Tavneet. 2011. "Selection and Comparative Advantage in Technology Adoption." *Econometrica* 79(1):159–209.
- Thompson, Owen. 2018. "Head Start's Long-Run Impact: Evidence from the Program's Introduction." *Journal of Human Resources* 53(4):1100–39.
- Verdier, Valentin, and Andrew Castro. 2019. "Average Treatment Effects for Stayers with Correlated Random Coefficient Models of Panel Data." Unpublished.
- Wiswall, Matthew. 2013. "The Dynamics of Teacher Quality." *Journal of Public Economics* 100:61–78.
- Wooldridge, Jeffrey M. 2005. "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models." *Review of Economics and Statistics* 87(2):385–90.
- . 2019. "Correlated Random Effects Models with Unbalanced Panels." *Journal of Econometrics* 211(1):137–50.
- 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):910–21.