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

Douglas L. Miller*
Cornell University
and NBER

Na'ama Shenhav†
Dartmouth College

Michel Z. Grosz‡
Abt Associates

April 8, 2019§

Abstract

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

*Doug Miller, Policy Analysis and Management, Cornell University, Email: dlm336@cornell.edu

†Na'ama Shenhav, Department of Economics, Dartmouth College, E-mail: naama.shenhav@dartmouth.edu

‡Michel Grosz, Abt Associates, E-mail: michel_grosz@abtassoc.com

§We would like to thank Colin Cameron, Liz Cascio, Janet Currie, Hilary Hoynes, Pat Kline, Erzo Luttmer, Jordan Matsudaira, Zhuan Pei, Maya Rossin-Slater, Doug Staiger, Chris Walters, and participants at the AEA Meetings, Cornell, Dartmouth, CSWEP CEMENT Workshop, Mcgill University, NBER Labor/Children's Summer Institute, Northwestern, SEA Meetings, SOLE, Syracuse/Cornell Summer Workshop in Education and Social Policy, UC Merced, and the War on Poverty Conference at the University of Michigan. We are also grateful to Alex Magnuson and Wenrui Huang for excellent research assistance.

1 Introduction

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

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

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

[1] If there are no other control variables that vary within a group, then switchers provide all the identifying variation. In the presence of other control variables which themselves vary within a group, then there may be variation among non-switchers "net of controls". We focus on cases where this phenomenon is small in magnitude.

[2] See Bound and Solon (1999).

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

the 5,355 children in the sample with siblings, only 1,098 children reside in switcher households. Second, the loss of sample variation is systematically related to observables. The likelihood of being a switcher - and thus included in the FFE estimation - increases with the probability of treatment (over the support 0 to 0.5) and with the number of units per group (children in a family). Third, since these factors vary across subgroups, SI does as well. The FFE identifying sample misses 93% of the sibling sample for white children, but only 62% of the sample for black children. Fourth, as a result, switchers are not representative of the overall sample along many dimensions. The most striking imbalance is along family size, but differences in income and parental education are also apparent.

Next, we show that under heterogeneous treatment effects, SI can meaningfully change the estimated treatment effect. The consequence of this is that the FFE estimate is no longer representative of the sample Average Treatment Effect (ATE), let alone the treatment effect for a policy-relevant population, such as program participants. This also implies that the difference between the OLS estimate and FE estimate can no longer be interpreted as solely reflecting OLS bias. Because we are more likely to have non-switching FE groups when they are defined over a smaller groupings of observations, the impact of SI may be stronger in this case. In some settings this means that standard FE methods may lead to a tradeoff between external validity and bias.

To address the change in estimated treatment effect due to SI, we develop a novel approach to reweight the FE estimates to obtain the ATE of policy-relevant "target" populations. Drawing on propensity score methods, we estimate an index of the likelihood of being in the target population (e.g. program participants) and in the switcher sample using a multinomial logit model. We then use the ratio of these probabilities to upweight observations that are under-represented in the identifying sample relative to the population of interest. This approach follows the spirit of the literature that examines extrapolating experimental results to other populations (Stuart et al., 2011; Andrews and Oster, 2018). We implement this using weighted least squares ("in-regression weighting") and post-regression weighting. This extrapolation relies on a conditional independence assumption: conditional on covariates, the treatment effect is assumed to be independent of whether an individual is in a switching group or not, and whether they are in a target population or not.[4]

We demonstrate the effectiveness of our reweighting using Monte Carlo simulations in a setting with naturally occurring SI. We show that when treatment effects are a function of observable variables, these methods are effective at reducing or eliminating bias. As an auxiliary finding, we show that for binary outcomes both a linear probability model with FE and conditional logit model can be used to obtain unbiased ATE for switchers, as long as treatment effects are properly extrapolated from the non-linear model.

Based on these findings we propose new standards for practice when presenting results using FE research designs: (1) clearly show not only total sample size, but additionally sample size when

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

limited to switcher families (and also for relevant subsamples within the data); (2) show the balance of covariates across switcher and non-switcher families (e.g. Table 2); (3) Reweight FFE estimates for a representative population (e.g. Table 5). We are not the first to use the more rigorous reporting standards in (1) and (2), but in our survey of the literature, the vast majority of papers do not discuss either of these issues.[5]

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

Using data from the PSID and the Children of the National Longitudinal Study of Youth (CNLSY) (as in GTC and Deming (2009)), we newly document that, across multiple human capital measures, there are patterns consistent with greater returns to Head Start in larger families. This might result from the fact that parental time investment in children's human capital is spread more thinly in larger families, which in turn could lead to greater returns to alternative investments (such as Head Start) in these families.[7] Since these families are upweighted in FFE models, then, it is intuitive that the FFE estimate is likely to be upward-biased.

Conforming with this intuition, our multivariate reweighting strategy gives smaller estimates of the impact of Head Start relative to FFE. We illustrate this first using the largest sample of siblings used to investigate this question, three times as large as the analysis in GTC. The FFE estimate in the PSID suggests that Head Start leads to a statistically significant 12 p.p. increase in attendance of some college. Using our reweighting methods, however, we find more modest and less-precisely-estimated benefits of the program. We estimate that Head Start leads to a 3.1 percentage point (p.p.) increase in the likelihood of attending some college for Head Start participants (s.e. = 6.1 p.p.), and a 7.1 p.p. (se=6.0 p.p.) increase for the Head Start eligible population. The ATE for Head Start participants estimate is 74% smaller than the FFE estimate, a difference which is significant at the 5 percent level. It is also 89% smaller than prior estimated effects on college-going for this population (GTC), 22% to 89% smaller than unadjusted estimates for all participants from

[5]Important exceptions include Finkelstein, Gentzkow and Williams (2016), who include a substantive discussion and examination of external validity concerns, as well as Currie and Rossin-Slater (2013). GTC report the number of identifying observations in aggregate (not for subsamples), and Deming (2009) reports the number of identifying observations for all samples.

[6]See Gibbs, Ludwig and Miller (2013) for an overview of Head Start, including programmatic details.

[7]In Section 6 we examine whether this heterogeneity by family size is likely driven by other covariates, or by larger families having longer sibling cohort spans. We do not find evidence that this is the case. Instead it appears that there is something important about family size per se.

other FFE studies (Bauer and Schanzenbach, 2016; Deming, 2009), and 42% smaller than estimates from the county roll-out of Head Start (Bailey, Sun and Timpe, 2018), although the lower end of the confidence intervals for the latter estimates include our ATE.

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

The core contributions of this paper are to show the importance of heterogeneity in treatment effects across switching and non-switching groups; and to provide a reweighting estimator that allows for the recovery of ATE for policy-relevant populations. This is different from strategies that employ reweighting for internal validity, such as traditional propensity score estimation methods. The FE estimation strategy that forms our focus is also distinct from traditional difference-in-difference strategies, or related strategies that employ both group and time fixed effects. In those strategies, there are groups that have nominally unchanging treatment. However, once group and time fixed effects are partialed out it turns out that all groups contribute toward identifying variation.[8]

Methodologically, we build on the literature that documents the difference between "what you want" and "what you get" from standard estimators. Included among these are studies on extrapolation from experimental estimates to the population ATE (Stuart et al., 2011; Andrews and Oster, 2018). We discuss this literature in detail in Section 2.1. Closest to this paper, Gibbons, Suarez and Urbancic (2018) derive the ATE bias of the FE estimator when treatment effects and the conditional variance of treatment vary across groups. A non-innocuous assumption of that paper is that the conditional variance of treatment is positive for all groups. We illustrate the bias of the estimator when this assumption is relaxed. Further, we show how the violation of this assumption is tied to characteristics of households, and that it has a meaningful impact on estimates. We also provide a flexible method for deriving treatment effects for populations of interest other than the ATE, including the TOT. Different than earlier papers, we consider extrapolation when the identifying sample is not necessarily a subset of the population of interest, and when identifying variation comes from within-group comparisons.

We present our results in the context of FFE and Head Start, but they apply to any panel fixed effects model, with special relevance for those with short panels and lumpy (e.g. binary) treatment variables. For example, consider the set of studies that examine the effect of peers in the classroom, such as in Carrell and Hoekstra (2010) who examine the effect of having a peer exposed to domestic

[8]Our approach can be applied to event study specifications identified off of switching groups, when the policy-relevant population of interest also includes non-switching groups.

violence (DV), using school-grade FEs. Since DV is relatively infrequent, over a short period, some school-grades will never have such a student. Further, there may be some school-grades that always have one student with DV exposure. If DV is correlated with factors that could mediate the effects of DV, such as school income, teacher experience, or the presence of counseling services services, the FE coefficient will incorrectly weight these heterogeneous treatment effects.[9]

Finally, we contribute to a growing body of work investigating the long term effects of Head Start using quasi-experimental methods (Ludwig and Miller, 2007; Carneiro and Ginja, 2014; Thompson, 2017; Bauer and Schanzenbach, 2016; Johnson and Jackson, 2017; Bailey, Sun and Timpe, 2018, in addition to the FFE papers above). These studies typically present LATE or ITT estimates, and find improvements in childhood health and increases in educational attainment among earlier cohorts of participants, and reductions in behavioral problems, health problems, and obesity in later childhood and early adolescents for later cohorts of participants. Relative to most of these studies, we evaluate the effect of Head Start on longer-run outcomes through longitudinal tracking of individuals and also adjust estimates using covariate re-weighting to get closer to the ATE for Head Start participants. We show that incorporating this adjustment lowers the estimated the long term effect of Head Start.

2 Literature

2.1 Reweighting for External Validity

This paper builds on prior works that decompose the implicit weighting schemes of standard OLS and FE estimators. Angrist (1998) and Angrist and Pischke (2009, Section 3.3.1) show that the OLS estimator with a binary independent variable can be represented as a weighted average of covariate-specific treatment-control comparisons. The OLS weights do not necessarily recover the ATE or TOT. Sloczynski (2017) shows that OLS can be represented as a weighted average of the treatment effect estimate for treated individuals, and the treatment effect estimate for untreated individuals, with the somewhat surprising result that the weights are in *inverse* proportion to each group's share of the sample.

In a fixed effects context, Gibbons, Suarez and Urbancic (2018) show that panel FE estimates can be represented as a weighted average of the group-specific estimated treatment effects, where the weights are connected to the inverse of the conditional variance of treatment. Panel FE estimates therefore do not usually recover the ATE or TOT. They provide two solutions to obtain the ATE: (1) weighted least squares, where the weights "undo" this variance reweighting; (2) a post-regression averaging of the estimated group-specific treatment effects. This approach has two important

[9] One alternative setting where our results may be relevant is the recent set of studies that identify the effects of the environmental shocks on health and human capital. Since these events can be infrequent, over a short period, some geographic groups will not have variation. Similar intuition can be applied to studies that identify geographic effects from the behavior of movers. In this context, one potential correlate of SI is occupation, since some occupations are less mobile than others (e.g. lawyers, due to licensing).

limitations, however. First, it only recovers the ATE if all groups have variation in the treatment variable. Second, if the sample is not representative of the policy-relevant population (e.g. likely participants), the ATE for the sample may not be of great interest. In a related vein, the difference-in-differences estimator provides uneven weighting of treatment effects when there is variation in the timing of treatment, which can also cause the estimate to deviate from the ATE of interest, and may require reweighting (see, e.g. Goodman-Bacon, 2018; Borusyak and Jaravel, 2017; Callaway and Sant'Anna, 2018).

Our paper is also connected to the literature on IV and local average treatment effects (Angrist, Imbens and Rubin, 1996; Imbens and Angrist, 1994), which centers on the issue that shifting to instrumental variables variation changes not only the *type* of variation but also *over whom* this variation is being averaged. IV changes the relative weights among the observations, in proportion to the variation they bring in the instrument. When treatment effects vary across individuals, OLS, IV, and FE can each provide different estimates, even when each of the approaches is valid in terms of exogenous variation (Lochner and Moretti, 2015; Loken, Mogstad and Wiswall, 2012). Lochner and Moretti (2015) propose reweighting OLS estimates aiming to put OLS and IV on even footing. Angrist and Fernandez-Val (2013) show that differing choices of instruments may cause estimates to vary; they offer covariate-reweighted LATE estimates that aim to make these estimates more comparable.

FE models with short panels or lumpy (e.g. binary) treatments, which are our focus, raise many issues that are similar to those raised in Gibbons, Suarez and Urbancic (2018) and Angrist and Fernandez-Val (2013). There are also some new issues that arise. Most importantly, for many groups, there is no within-group variation in the covariate of interest. This is especially likely to happen when the covariate of interest is binary. As such, researchers need to work with the limited subset of groups that experience within-group variation in treatment. We propose simple diagnostic checks for exploring this issue, and suggest weighting schemes to address the unique aspects of our setting and which aim to recover the average treatment effects over populations of interest.

Our proposed solution for FE models is closely related to the literature focused on the external validity of randomized experiments and on extrapolating from an experimental population to other populations. Stuart et al. (2011) examine experiments in which treatment effects depend on observable covariates and propose diagnostic measures for assessing the ability to extrapolate. Andrews and Oster (2018) show that there are in principle infeasible weights that could be used to extrapolate from experimental effects to the population ATE, and that these weights depend in part on the *unobservable* determinants of the selection into identification process. They relate the infeasible extrapolation bias to the observable "participation on observables" bias. Correcting for the latter bias will allow for extrapolation to the extent that (1) selection into identification is largely a function of observables, or (2) treatment effect heterogeneity is largely a function of observables (or both). As we discuss in Section 4, selection into identification from FE may be a setting where explanatory factors are more likely to be observable. This is because selection is

driven by within-group variation in treatment.

2.2 A Survey of FFE Applications

Since our application focuses on a FFE model, we focus on applications of this particular method in the literature. This focus will lead us to undercount the prevalence of FE more broadly, but provides an unambiguous example of a short-panel setting which is susceptible to SI concerns. We surveyed publications from January 2000 to May 2017 in 11 leading journals that publish applied microeconomics articles. We include all studies that use family fixed effects as a primary or secondary strategy.[10]

Our literature review yields 58 papers. We provide descriptive statistics of these articles in Table 1. The first panel tabulates the frequency of binary treatments and binary outcomes across the sample of papers, the focus of our methodological insights. These forms of variables appear frequently. Nearly two-thirds (37) of the papers have a binary treatment of interest and 25 have a binary outcome. The second and third panels show the varied topics that appear in the sample, spanning health, public, education, and labor fields.

The final panel of the table summarizes the distribution of sample sizes used with FFE. The samples are frequently not limited to families with variation in the treatment variable; therefore, the sample size in the table is an upper bound on the number of observations used for identification. The median number of sibling observations is 6,792, or roughly 85% of the sample in our analysis. It is important to note that there is a high variance in sample size across samples, indicating that there is not a threshold for FFE analyses. The bottom 25% of papers have fewer than 1,200 observations, while the top 25% have over 175,000 sibling observations.

Appendix Figure B.1 illustrates the salience 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 181 citations per article (Google Scholar citations as of May 2017).

3 Selection into Identification

We examine two methodological issues that arise from the FE research design: *(i)* reduction in identifying variation; *(ii)* a change in the composition of the identifying sample. We illustrate these issues in the context of a FFE example using micro data, which we draw from our analysis of the impacts of Head Start in the second half of the paper. Therefore, in this section, we will use the

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

term "families" to refer to cross-section groups, and "Head Start" to refer to a binary treatment variable, but the intuition can extend more broadly.

To solidify ideas, we provide an outline of the data we use in the example - for more detail, see the descriptions of the PSID data in Sections 6.1 and 6.2. The sample consists of 2986 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. We include many control variables, including a dummy for other preschool attendance and parental and early-childhood socioeconomic circumstances. The coefficient on Head Start in a cross-section regression is 0.049 (s.e. = 0.044). When mother fixed effects are added, the coefficient becomes 0.120 (s.e. = 0.053). This result indicates that the impact of Head Start participation on college attendance is meaningful in magnitude, and statistically significantly different from zero.

3.1 Empirical Relevance

In the FFE setting, treatment effects are identified from switcher families. This implies that the ex-post effective number of observations — that is, those that contribute to identifying the treatment effect — may be quite small and not representative of a population of interest.

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 within a family g, $HeadStart_i$-$\overline{HeadStart_{g(i)}}$, against the within-family deviation in attainment of some college for the white sample, $AnyCollege_i$-$\overline{AnyCollege_{g(i)}}$.[11] The size of each symbol is weighted by the number of individuals. 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. Moreover, there are many additional families with no within-family deviation in Head Start but some variation in college attendance, as illustrated by the vertical alignment of large bubbles. When we remove observations for families with no variation in Head Start, who are centered on the y-axis, the number of observations drops substantially from 4761 to 213.

This reduction in identifying observations could result in a selected sample if switching is correlated with family characteristics. To gain intuition about which variables might determine switching – and hence, influence the reduction in observations – we build a simple model of the Head Start (HS) participation decision within families. We assume that the probability of attending Head Start is a constant, π, and independent across siblings in a family, such that the likelihood of attending is a lottery within families. The probability of switching, is then a function of π and family size, n_g:

[11] 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 are possible because not all families have just two children. Values of -0.5 and 0.5 are also possible in families with more than 2 children, if equal numbers of children participated as did not. 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.

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

Appendix Figure B.2 graphs the relationship between $P(HSSwitchingFamily)$ and π for families with 1, 2, 3, 4, or 5 children according to this formula. It shows that the probability of switching has an inverse-U-shaped relationship with π, implying that the reduction in observations will be larger for populations with very high and very low π, and smaller when π is close to 0.5. And for a given level of π, the likelihood of being in a switching family is increasing with family size.

The markers in Figure 2 identify the observed probability of attending Head Start and of being in a switching family for each family size-group by black/white race and by whether the mom has some college or not in the PSID. As in the stylized model, the likelihood of switching is increasing with family size for each of these subgroups.[12] Appendix Table B.1 shows that this pattern is driven by a much larger incidence of no Head Start participation among smaller families,[13] which, in turn, reduces the likelihood of switching. 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% of the sibling sample for whites, and 38% of the sibling sample for blacks. Note that while we are focusing on race and maternal education, this notion can be generalized to any other family characteristic, such as SES, that determine π.

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

3.1.1 Selection into Identification Driven by Many Variables

Since SI is likely to affect the balance of characteristics other than family size (such as those associated with the probability of Head Start), we now examine a large number of observable characteristics of switcher families and non-switcher families. Panel A of Table 2 indicates that in addition to having a larger family size, children in switcher families tend to have parents with significantly less education than children in non-switcher families (column 3). These differences in parental education are significant even in a regression framework where we control for differences in family size and the other covariates in the table, though only at the 10 percent level (columns

[12] As expected, the observed relationship with family size is quantitatively smaller than predicted by the model, because Head Start participation is correlated across children within a family.

[13] For example, 78% of 2-child families have no Head Start participants, compared with 48% of families with 5 or more children.

4 and 5). We also see that the family income during preschool of children in switcher families is significantly lower than non-switcher families overall (some of which may have too high of income to ever qualify for Head Start).[14] These patterns are consistent with switching increasing with the probability of Head Start participation, depicted in Figure 2.

In Panel B of Table 2, we summarize the differences between switchers and non-switchers by examining how much overlap there is in the characteristics of switchers and non-switchers with (1) the switcher population and (2) the Head Start participant population.[15] We do so by estimating the propensity for each individual to appear in (1) and (2) using a multinomial logit. We describe this procedure in detail in Section 4.3 below.[16] Among both switchers and non-switchers, the predicted probability of being a switcher is larger than the predicted probability of being a Head Start participant. More importantly, the ratio of these probabilities is larger for switchers than for non-switchers. As a benchmark, Stuart et al. (2011) suggest that a 0.1 to 0.25 SD difference in propensity scores between the experimental and non-experimental population may be too large to rely on extrapolation without further adjustments. The mean of this ratio for switchers in our sample is 0.3 SD higher than for non-switchers. This suggests that we may need to account for the differences in the covariates across the two populations to get an acceptable extrapolation.

3.2 Consequences for Estimation

We now decompose the FE estimator to understand how selection into identification alters the weighting of marginal effects and how this compares to the weighting of the OLS estimator. The usual interpretation of the difference between OLS and FE estimators is that FE removes bias. We show that the change in weighting also contributes to the wedge between the estimators and, distinct from prior work, that the change in weights from SI is a particularly important concern for interpretation of FE estimates.

Under homogeneous treatment effects, SI has no effect on expected bias in estimation; there is loss of precision that accompanies the overall reduction in sample size. The more interesting case is when treatment effects are heterogeneous.[17] A useful starting point is to consider the case where heterogeneity in the effect of a binary treatment, D_i, such as Head Start. We suppose that there is a discrete covariate that varies at the group level, Z_g, that determines this heterogeneity. That is, there is a different treatment effect for each value of z, δ_z. For now, we concentrate on heterogeneity along family size in a FFE model, such that δ_z is the treatment effect for a family with size $Z_g = z$. In Section 4.3, we allow for multivariate heterogeneity with multiple continuous covariates. Using

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

[15] Results are similar when we consider siblings as an alternative target population.

[16] Specifically, this table shows the mean and standard deviation of the (inverse of the) post-regression weights that we construct for the Head Start participant target population.

[17] In the context of experimental designs, Andrews and Oster (2018) show that extrapolating from the estimation sample to a target population has greater bias when: (1) the underlying characteristics of the estimation and target population are different, (2) there is greater variability in the treatment effect, and (3) there is higher correlation between selection into the estimation sample – "participation decisions," in their phrasing – and treatment effects.

the regression decomposition formula (Angrist, 1998; Angrist and Pischke, 2009),[18] the treatment effect estimated from the sample of multi-unit groups ($n_g \geq 2$), e.g. siblings, is:

$$\delta_{OLS} = \sum_z \delta_{z,OLS} \cdot \omega_{z,OLS} \tag{2}$$

where

$$\omega_{z,OLS} = \frac{(\sigma^2_{D_i|n_g\geq 2, Z_g=z}) \cdot Pr(Z_g = z|n_g \geq 2)}{\sum_{z'}(\sigma^2_{D_i|n_g\geq 2, Z_g=z'}) \cdot Pr(Z_g = z'|n_g \geq 2)}$$

$\delta_{z,OLS}$ is the OLS estimate of the treatment effect for groups with $Z_g = z$, and $\sigma^2_{D_i|n_g\geq 2, Z_f=z}$ is the conditional variance of treatment among the sample of multi-unit groups with $Z_g = z$, net of other control variables. Let S_g be a binary variable that denotes a group's switching status, which takes on a value of 0 when $Var(D_i|i \in g(i)) = 0$, and 1 otherwise. The fixed effect estimator for the sample of multi-unit groups, e.g. siblings, can be written as:

$$\delta_{FE} = \sum_z \delta_{z,FE} \cdot \omega_{z,FE} \tag{3}$$

where

$$\omega_{z,FE} = \frac{(\sigma^2_{D_i|within, Z_g=z}) \cdot Pr(Z_g = z|S_g = 1}{\sum_{z'}(\sigma^2_{D_i|within, Z_g=z'}) \cdot Pr(Z_g = z'|S_g = 1)}$$

and $\delta_{z,FE}$ is the FE estimate of the treatment effect for groups with $Z_g = z$, $\sigma^2_{D_i|within, Z_g=z}$ is the conditional variance of treatment among the sample for groups with $Z_g = z$, net of family fixed effects and other control variables.

Moving from OLS to FE, the δ's change and also the ω's change. The change in the δ is how we usually interpret the move from OLS to FE. But the full change also incorporates the different weightings of different values of Z_g. If the OLS sample and the FE sample overlap in the covariates, we can decompose the difference between OLS and FE to identify how much is caused by the change in weights, ω_z, and how much is driven by the change in identification, δ_z, as:

$$\delta_{FE} - \delta_{OLS} = \sum_z \underbrace{[\omega_{z,FE} - \omega_{z,OLS}] \cdot \delta_{z,FE}}_{\text{Impact of } \Delta \text{ weighting}} - \underbrace{\omega_{z,OLS} \cdot (\delta_{z,OLS} - \delta_{z,FE})}_{\text{OLS Bias}} \tag{4}$$

The impact of SI is captured in the first element of Equation 4, which gives the disparity between the FE estimator for the sibling population and the switching population. This incorporates differences in the probability of each family size appearing in FE and OLS as well as the differences in the conditional variance.[19] Since switchers are typically not a population of interest, this raises

[18] See equation 3.3.7 on page 75 of Angrist and Pischke (2009).

[19] Our decomposition is a special case of Equation 13 in Loken, Mogstad and Wiswall (2012), which provides a

concerns for the external validity of the FE estimator.

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

We use data from our empirical example to illustrate the change in the components of ω_z across OLS and FE. We consider situations where OLS is estimated on all individuals or only siblings. In the interest of brevity, we include the details in Appendix Table B.2, and summarize our main findings here. Consistent with the results above, the proportion of 5+-child families in the switching sample is roughly twice the proportion in the overall sample, while the share of 3 and 4-child families is roughly constant. This will tend to upweight the coefficients of 5+-child families in the regression.

For every family size the variance in Head Start is higher, roughly double, in the switching sample relative to the sibling sample. The average effect is unlikely to be affected by this, though, since the increase is relatively similar across family sizes . Thus, in our setting, the change in the conditional variance across OLS and FE plays a minor role, while the change in the distribution of family sizes is substantial.

We then calculate $\omega_{z,OLS}$ and $\omega_{z,FE}$, which combine these two inputs. Going from the sibling sample to the switchers sample, $\omega_{2-child}$ declines by over 25% and $\omega_{3-child}$ declines by 15%. On the other hand, the $\omega_{5-child}$ nearly doubles from 0.134 to 0.243, and the $\omega_{4-child}$ families increases by over 25%.

We also see that δ_z varies in our applications. The first two columns of Panel A of Table 3 shows the estimated effects of Head Start on the likelihood of completing some college by the number of children in a family for our illustrative sample of PSID white adults. We show the results with and without family fixed effects. In both specifications, the effect of Head Start is significantly higher among white children in families with 5 or more children and, once fixed effects are added, the effect of Head Start is monotonically increasing with the number of children in a family.

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

Like with SI, the larger Head Start effects we document for big families is not unique to the PSID. Columns (3) to (5) of Table 3 show the CNLSY FFE estimated effects of Head Start by family size for idleness, having a learning disability, and being in poor health.[20] For each of these

general formula for the comparison of OLS and FE estimators.

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

outcomes, the impact of Head Start for 5+ child families is at least twice as large as the impact for 2 or 3 child families. For high school graduation, we also see a large impact for 4-child families, roughly double the impact for 2 and 3 child families. This implies that we should expect an increase in the coefficient going from OLS to fixed effects due to the *change in weighting* across the identifying samples, even without a change in the source of identification.

3.3 Nonlinear Functional Form

Throughout, we use the linear probability model (LPM) as the primary specification for binary outcomes, as is almost universally done in our review of FFE papers. In Appendix Section D, we examine whether SI concerns are sensitive to functional form modeling assumptions. One reason this may make a difference is that conditional or fixed effect logit and probit models use less variation relative to LPM. With these models, for any families that have no variation in outcomes, i.e "all successes" or "all failures", the fixed effect parameters will be driven to +/- infinity, and these families will be dropped from estimation. This leaves only "double switchers": families with variation in both the outcome variable and the treatment variable. Monte Carlo exercises reveal that, in general, the bias of LPM and conditional logit is similar, and that the reweighting we propose is equivalently effective at reducing bias for LPM and conditional logit.

4 Solutions

We propose two methods to flexibly obtain the ATE for populations of interest, which we refer to as "target" populations. Commonly, the target population in applied work is the ATE for a nationally representative sample, which may be a reasonable starting place for most researchers. For some treatments, like means-tested programs, one might be interested in the ATE for eligible individuals or for participants.

4.1 Assumptions

The methods rely on several key assumptions. First, we assume that the FE estimator is unbiased at the group level. That is, we rely on the traditional FE assumption that conditional on the fixed effects and control variables, treatment is as good as randomly assigned with regard to potential outcomes.

Our second main assumption is a variant of the traditional conditional independence assumption (CIA). We assume that conditional on observables, the true treatment effect is independent of a group's switching or target status:

$$E\left[\delta_g \perp (S_g, target_g) \,|X_g\right]$$

This assumption is related to the ones employed in the literature on extrapolation from exper-

iments. In that literature the CIA requires that the participation decision be independent from the treatment effect, conditional on observables. This type of assumption is also employed with other reweighting methods (e.g. Angrist and Fernandez-Val, 2010). The CIA relies on selection into identification being driven by observable covariates (and random chance).

In the Head Start context, the key determinants of variation in participation across children, captured in Figure 2, are family size and the underlying probability of Head Start participation. 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. Family size is observable, and observable covariates, such as income, can take us a long way in predicting program participation. We argue that similar forces may predict selection into identification in other settings as well, particularly when there are clear, observable requirements for participation in treatment (e.g means-tested programs).

Because we employ a CIA, we also require common support in X's, or at least in a propensity score. This can rule out use of certain covariates in the conditioning set. For example, we can never have singletons in the switching sample, so we need to rule this out as a conditioning covariate.

Finally, we note that our CIA has one potentially testable implication: among the switchers, the average treatment effect is the same (conditional on covariates) for those in the target population as those not in the target population: $E\left[\delta_g | X_g, (S_g = 1, target_g = 1)\right] = E\left[\delta_g | X_g, (S_g = 1, target_g = 0)\right]$. Whether this can be tested in practice will depend on whether the definition of the target population allows for a partion among the switchers into two groups. For example, if the target population is "siblings" or "everyone", then there will be no non-target individuals among the switchers.

4.2 Univariate Heterogeneity

If the source of heterogeneity in estimates is a single, discrete covariate, such as family size, a simple solution is to reweight the family-size specific estimates to obtain the ATE for a representative target population. If the target population are siblings, this is given by

$$\delta_{ATE}^{tg=sibs} = \sum_z s_z \cdot \delta_z \tag{5}$$

where s_z is the share of the target population – in this case, siblings –that have characteristic z, and δ_z is the treatment effect for groups with characteristic z. This approach is similar to the "Late-Reweight" concept in Angrist and Fernandez-Val (2013).[21]

In a similar vein, this expression can be adapted to instead use OLS weights, $\omega_{z,OLS}$, which allows us to measure the change from OLS to FE attributable only to the change in identification:

$$\delta_{FE,\omega_{OLS}} = \sum_z \omega_{z,OLS} \cdot \delta_{z,FE} \tag{6}$$

[21] See Equation 9 of Angrist and Fernandez-Val (2013).

Reweighting the $=\delta_z$ estimates in this manner has a meaningful impact on the coefficients we estimate our data example. We use the coefficients from column (2) of Table 3 and the sibling weights we developed earlier[22]. This produces a coefficient of 0.083, which is included in the second column of Panel B of Table 3. This implies that (under the assumption of univariate heterogeneity) the FFE estimate is 50% higher than it would be if we were able to estimate FFE with the OLS sibling population weights. This reinforces our intuition that changes in weighting can have substantial influence on the estimated coefficient.

As an intermediate step, we also quantify how much of the difference between OLS and FE is attributable to (1) changing the *weighting* across different effect sizes, holding constant the cross-sectional identification and (2) moving from *"bad variation"* (between families) to *"good variation"* (within families), holding constant the FE weights. Taking the OLS family-size-specific coefficients from column (1) of Table 3 and reweighting by the fixed-effects regression weights, $\omega_{z,FE}$, we obtain a weighted coefficient of 0.069, shown in the bottom row of Table 3. This represents the effect of Head Start on the switcher population using cross-sectional variation.[23] Recall that the OLS coefficient on Head Start is 0.049 (se=0.044), and the fixed effects coefficient is 0.120 (se=0.053). So approximately 1/3 of the change from OLS to FE ($\frac{0.069-0.049}{0.12-0.049}$) is driven by the change in family size weights; with the other 2/3 driven by change in identifying variation.

4.3 Extrapolating from Identifying to Target Populations: Treatment Effect Heterogeneity based on Several Covariates

For the case where treatment heterogeneity maps onto *multiple* covariates we develop a more general reweighting technique for extrapolating to a target population. Differently from the univariate case, we now allow treatment effects to vary for each group, rather than varying only by discrete values of Z_g.

In this setting, the average treatment effect for a target population is given by

$$\delta_{ATE}^{tg=target} = \sum_{g \in switcher} s_{g,target} \cdot \delta_g.$$

where $s_{g,target}$ is share of the target population with observable characteristics matching group g, who is in the switcher sample. Under the maintained assumption that family-specific treatment effects are unbiased and treatment effect heterogeneity is a function of observable variables X_g, we can use group-specific treatment effects, δ_g, from the switcher sample to construct $\delta_{ATE}^{tg=target}$. δ_g can be obtained from a regression of the outcome, y_{ig}, on the interaction between D_i and group-specific dummies.

We construct weights to make the families in the switcher sample representative of the tar-

[22] See Panel C of Appendix Table B.2.

[23] We verify that these weights work as intended, weighting the fixed effect coefficients by the weights for the switcher sample, obtaining 0.123 (very close to the FE estimate of 0.12).

get population. These weights are $s_g^{sw \to tg} = \frac{Pr(i \in TargetPopulation|X_g)}{Pr(i \in SwitcherSample|X_g)} \times Pr(g|SwitcherSample)$.[24]. Intuitively, this will upweight groups with *characteristics* that are underrepresented among switching groups. We provide a simple derivation to support this intuition in Appendix A. In the case where all switchers are a subset of the target population, the weights are $\frac{1}{Pr(i \in SwitcherSample|X_g)} \times Pr(g|SwitcherSample)$ This can be estimated by a simple logit or probit model.

It is not always the case that switchers are a subset of the target population; an observation might be in one or both (or neither) of the target population and the switching sample. This results in four possible categories that describe the switcher and target status of a particular observation. To allow for this range of possibilites, we use a multinomial logit model to estimate the probability of each outcome, where the outcomes are indicators for the four possible combinations of being in the switcher sample (or not) and being in the target sample (or not). The numerator for the weight, $Pr\left(i \in TargetPopulation \mid X_g\right)$, is then constructed as the sum of the probability of being in the target population and not being a switcher and the probability of being in the target population and being a switcher. The denominator is the probability of being in the switcher sample, and is constructed as the sum of the probability of being in the target population and being a switcher and the probability of being in the non-target population and being a switcher.

A feature of this approach is that we are able to extrapolate estimates to populations that include groups that have no switchers, such as single-unit groups (e.g. one child families). The extrapolation is based on the assumption that, conditional on the covariates, the treatment effect is the same for switchers, non-switching multi-unit groups, and single-unit groups. To implement the extrapolation we cannot use group-size dummies or other covariates that could perfectly predict not being in the switcher sample. Doing so would result in a violation of the common support assumption. Thus, in our setting we control for family size with a polynomial.

We can construct our ATE estimate for the target population in one of two ways. The first is a two-step "post-regression weighting" of δ_g:

$$\widehat{\delta_{ATE,2step}^{tg=target}} = \sum_{g \in switcher} \widetilde{s_g^{sw \to tg}} \cdot \widehat{\delta_g} \tag{7}$$

with $\widetilde{s_g^{sw \to tg}}$ the normalized weights, $\widetilde{s_g^{sw \to tg}} = \frac{s_g^{sw \to tg}}{\sum_{g' \in switcher} s_g^{sw \to tg}}$.

Under standard cluster-robust assumptions, the $\hat{\delta_g}$ are independently distributed from one another, so we can obtain a cluster-robust variance estimate as[25]

$$\widehat{V}(\widehat{\delta_{ATE,2step}^{tg=target}}) = \sum_g (\widetilde{s_g^{sw \to tg}})^2 \cdot \left(\widehat{\delta_g} - \widehat{\delta_{ATE,2step}^{tg=target}}\right)^2 \tag{8}$$

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

[25] As Gibbons, Suarez and Urbancic (2018) note, we cannot estimate cluster-robust standard errors in the estimation step: there are fewer clusters than the sum of the count of fixed effects and covariates. However the standard cluster-robust assumptions imply that the δ_g are independent of one another. This enables the formula (8), which is an additional contribution of this paper.

A second approach is to obtain the ATE in a single step using "in-regression weights." For this, we need to adjust for the fact that the FE estimator uses weights ω_{FE} rather than population shares. We address this by incorporating inverse conditional variance weights, as $v_g = \left(V\left(\ddot{D}_i \mid groupID = g; X_f\right)\right)^{-1/2}$, where $\ddot{D}$ is the residualized measure of treatment after partialing out the group fixed effects and other covariates (Gibbons, Suarez and Urbancic, 2018). Then, the ATE can be estimated by $\delta^{\widehat{tg=target}}_{ATE,1step}$ from a one-step regression using $s_g^{\widetilde{sw \to tg}} \cdot v_g$ as regression weights. This more parsimonious approach may be more convenient for estimation. It also makes computation of cluster-robust standard errors straightforward.

4.4 Monte Carlo for LPM FE

We perform a Monte Carlo analysis to examine the properties of the LPM FE estimator and of our proposed reweighting estimators. The goal of this exercise to understand the bias and mean squared error of each of these estimators in settings where the true ATE is known, so we can gauge the tradeoffs between using one over the other. We examine three settings, each of which has a different assumption of heterogeneity in treatment: no heterogeneity; heterogeneity along a discrete covariate; treatment effects that vary with multiple group-level covariates. In each case, we assume that the researcher has some knowledge about the covariates that determine heterogeneity, which determines the covariates that are used to generate the propensity score.

4.4.1 DGP

We build our Monte Carlo around a data set that is designed to reflect the variability in our PSID Head Start application data set. In other words, we do not model SI, but rather use naturally occuring SI in the data. This allows us to report on the effectiveness of our reweighting procedure in a realistic context.

We begin by taking our original data, and running a linear probability model predicting "some college or more" educational attainment, with a set of family-level and individual-level covariates, including demographic variables, income during childhood, and parental education. From this model we construct a one-dimensional covariate X_g, which is a continuous probability that an individual completes some college at baseline (i.e. without treatment). All simulations start with this constructed variable X_g and the variable $HeadStart_{ig}$ from the original dataWe restrict the sample to those with $Pr[\widehat{College}_i = 1] = X_g \in [0, 1]$ at baseline, and for each DGP, we scale these baseline probabilities to ensure that inclusive of treatment the probability of some college is within the range $[0, 1]$.

We examine three specifications. Each specification varies the DGP that determines the treatment effect for each treated individual and, correspondingly, which control variables we use to generate the propensity scores for reweighting. For the first DGP, all treated observations experience a treatment effect of 8 p.p, which is added to the baseline X_g. We use the variable X_g to

generate propensity scores. For the second DGP, large families (with 4 or more siblings) have a constant treatment effect of 19.2 p.p., and small families (3 or fewer children) have zero treatment effect. We use a dummy variable for "large family" to generate propensity scores. For our third DGP, we allow the treatment effect heterogeneity to vary smoothly: the treatment effect in probability units is given by: $0.08 \cdot \left(1 - \frac{X_g - \bar{X}_g}{s.d.(X_g)}\right) \cdot \frac{1}{3}$, with $\bar{X}_g$ and $s.d.(X_g)$ the mean and standard deviation of X_g. This produces a treatment effect which is larger for lower-baseline-probability individuals, which varies smoothly across families, and which ranges from 0.01 to 0.15 for most of the population. We use X_g to generate propensity scores. We also consider an alternative model for this DGP in which the reweighting step uses a spline in X_g, with knots at the 5^{th}, 20^{th}, 50^{th}, 80^{th}, and 95^{th} percentiles of X_g.

4.4.2 Monte Carlo Results

We run 10,000 replications of our Monte Carlo simulation. In each replication, we keep track of (1) the true ATE for each target population of interest; (2) the FE estimate of the treatment effect, and (3) the reweighted regression estimate of the treatment effect for each target population.[26] The FE estimate is the same for all target populations. We consider four target populations. These include (1) individuals in Head Start switching families[27]; (2) all siblings (regardless of whether or not there is variation in Head Start in the family); (3) all individuals in the sample; and (4) all Head Start participants. We multiply all estimates by 1,000 for easier readability.

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

Panel B of Table 4 considers the second DGP, with zero treatment effect for small families, and large treatment effects for large (4+ children) families.It shows that FE is biased for the ATE for each of the target populations considered. The reweighting estimator is unbiased for each population considered. This improvement in bias over FE leads to much better mean squared error results for the reweighting estimator.[28]

Panel C of Table 4 examines the third DGP, with heterogeneous treatment effect that varies with X_g. For this DGP, the FE model does well for predicting the ATE for the switching sample, with only a small bias of -0.11 p.p. Its bias is also relatively small for the Head Start participant population. However, for all children, and for all siblings, the OLS model has a larger bias, on

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

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

[28] In results not reported, we have examined adding X_g as a covariate to the propensity score estimation stage. Inclusion of this covariates introduces a small amount of bias in the reweighting estimator for the "siblings" and "all" target groups. This underscores the fact that misspecification of the propensity score model can lead to an imperfect rebalancing. Even in that setting, however, the bias induced from this misspecification (-0.1 to -0.14 percentage points) is much smaller than that from the FE model (2.2 to 3.2 p.p.).

the order of 1.3 p.p. The regression reweighting estimator, which uses X_g in the propensity score estimation, has smaller bias 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_g variable, even after reweighting. When we reestimate the model including a spline in X_g in the propensity score estimation step this results in no detectable bias for any of the target groups.

4.4.3 Discussion

The results of this exercise show that that regression reweighting can greatly reduce bias (compared to FE) for the types of treatment effect heterogeneity we consider. Moreover, the reweighting estimator can be successfully targeted toward different target populations. Consistent with the conditioning on observables requirements of this estimator, its performance is best when it is given the appropriate covariates for the particular type of heterogeneity at work.

In Appendix Section D, we extend our Monte Carlo analysis to a nonlinear (logit) DGP, and additionally consider various logit FE estimators, including reweighted versions of these estimators. We find results similar to those reported here for linear estimators. Reweighted versions of (1) linear probability model, (2) logit with Mundlak controls, and (3) a 2-step logit model that we propose all perform equivalently well in our simulations.

5 Summary of Recommendations

Based on our findings, we propose new standards for practice when using FE or similar research designs. Above all, we recommend that researchers employ our diagnostic toolkit to quantify the role of changes in sample composition in explaining the gap between OLS and FE estimates, and concerns for external validity. First, analyses should report the switching sample size in addition to the total sample size, including for relevant subsamples of the data (e.g. whites and blacks). We found this reporting already in use in our survey of the literature, but very infrequently. 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 length of the panel (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. HOW MANY STUDIES CURRENTLY DO THIS? 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.

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. By allowing the data determine the predictors of switching, this serves as a more flexi-

ble means of diagnosing an external validity issue. We find that reweighting is helpful for recovering ATE's even when the outcome is binary and the underlying model is nonlinear. Nonetheless, since these methods can perform unevenly under some DGP's, we suggest testing for sensitivity of results and reporting a range of estimates where applicable.

6 Effects of Head Start

6.1 Data and Replication of GTC and Deming (2009)

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

6.1.1 PSID

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

We begin our analysis with a replication of GTC. The sample includes all black or white individuals born between 1966 and 1977, and excludes Hispanic individuals.[29] We provide a detailed description of our replication of GTC in Appendix C. Despite some minor differences, the two PSID samples are qualitatively similar. The summary statistics are often within a third of a standard deviation of each other. Moreover, the estimated effects of Head Start in this sample are similar to those estimated in GTC. We find large (23 p.p.) and significant effects of Head Start on the probability that whites attain some college, and large point estimates (9.3 p.p.) for high school graduation, though in our case these are not statistically significant. However, we do not find a meaningful reduction in the probability of committing a crime resulting from participation in Head Start.[30]

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

[29] This sample is intended to be representative of the Head Start population during the early years of the program. As pointed out in GTC, the number of immigrants was much smaller between the years 1960-1980, such that it is unlikely that many Hispanic immigrants would have benefited from Head Start.

[30] Moreover, 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 un-diagnosable) differences in specification and/or dataset construction.

In addition to modifications of the sample, we also expand the number of outcomes under analysis in order to gain a more extensive understanding of the channels by which Head Start affects children's lives. We follow the established practice of distilling the measures to summary indices to lessen problems with multiple hypothesis testing (see, e.g., Anderson, 2008; Kling, Liebman and Katz, 2007; Hoynes, Schanzenbach and Almond, 2016). We create four indices to capture economic and health outcomes observed for individuals at age 30 and 40. The "economic sufficiency index" includes measures of educational attainment, receipt of AFDC/TANF, food stamps, mean earnings, mean family income relative to the poverty threshold, the fraction of years with positive earnings, the fraction of years that the individual did not report an unemployment spell, and homeownership. The "good health index" summarizes the following component measures: non-smoking, report of good health, and negative of mean BMI.[31]

The process of creating each index follows the procedure described in Kling, Liebman and Katz (2007). In particular, we standardize each component of the index by subtracting the mean outcome for non-treated children, defined as children that did not attend any form of preschool, and then dividing the result by the standard deviation of the outcome for non-treated children. The summary index takes a mean of these standardized measures.[32] We also extract the first principal component of the standardized variables for "economic sufficiency" and for "good health". Later we use these as alternative outcome variables.

Appendix Table B.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/non-participants, and for the sample of individuals with siblings. We present the means of the analyzed outcomes in Appendix Table B.4.[33]

6.1.2 CNLSY

The CNLSY sample is identical to that used in Deming (2009).[34] The CNLSY is a longitudinal survey that follows the children born to the roughly 6,000 women that took part in the NLSY79 survey. The sample we use includes all children who were at least 4 years old by 1990.

[31] See Appendix Table B.5 for descriptive statistics of the inputs to the indices.

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

[33] Appendix Table B.5 includes summary statistics for the inputs to the summary indices. In each of the tables of summary statistics, the number of observations varies for each of the reported means. For example, the variable "ever booked or charged with a crime" was only collected in the 1995 wave, and so is only relevant for cohorts old enough to be at risk for that outcome by 1995. For parsimony, in Appendix Tables B.3 and B.4 we only report the number of individuals in the sample for whom we have information on their attendance of Head Start. A full accounting of the number of observations for each characteristic or outcome is available in Appendix Tables B.6, B.7, and B.8.

[34] This sample is downloaded from the replication files from the AEJ website.

6.2 Head Start Estimation

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

$$Y_{im} = \alpha + \beta_1 HeadStart_{im} + \beta_2 OtherPreSchool_{im} + X_{im}\gamma + \delta_m + \varepsilon_{im} \quad (9)$$

where Y_{im} represents a long-term outcome for individual i with mother m. $HeadStart_{im}$ indicates whether a child reports participation in the program, and $OtherPreSchool_{im}$ 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.[35] δ_m is a mother fixed effect which enables comparisons across siblings with a shared mother. When we perform post-regression reweighting, we include interactions between δ_m and $HeadStart_{im}$ to obtain family-specific estimates. The vector X_{im} includes a large number of controls for individual and family characteristics to absorb differences in personal and household characteristics which may be correlated with one's participation in Head Start and long term outcomes. These controls vary due to data availability across sources and specification used in earlier work, but fall into three broad categories: demographics, family background, and family economic circumstances during early childhood.[36]

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.[37] When Y_{im} is a binary variable, we estimate linear probability models as a main specification and check the sensitivity of our results to alternative models.

The coefficient of interest is β_1, the impact of Head Start on long term outcomes compared to no preschool. We generate propensity score weights to obtain the ATE for three target populations: (1) Head-Start-eligible individuals, based on family income between ages 2 and 5;[38] (2) all Head Start participants; and (3) all siblings.[39] For parsimony, we use a subset of the variables in Table

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

[36] For the PSID, these include: individual's year of birth, sex, race, and an indicator for being low birth weight, mother and father's years of education, an indicator for having a single mother at age 4, 4-knot splines in annual family income for each age 0, 1, and 2, a fourth spline based on average family income between ages 3 and 6, indicators for mother's employment status at ages 0, 1, and 2, and household size at age 4. This is a more expansive set of covariates relative to GTC, which did not include controls for maternal employment or family income prior to age 3. For the CNLSY, these include: health conditions before age 5, PPVT test score at age 3, measures of birth weight, measures of mother's health and health behaviors, mother's working behavior and income prior to age 4, indicator for being first born, participation in Medicaid, relative care, and indicators for early care types.

[37] 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.

[38] An individual is considered Head-Start-eligible if at any point between the ages of 2 and 5 her family income was below 150% of the poverty level. This is a more liberal definition that the official Head Start income threshold (100%), to account for our imperfect ability to observe reportable income.

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

2 to generate the propensity score.[40] We include results for the post-regression weighting method; results are qualitatively similar when we use in-regression weighting.[41]

6.2.1 Evidence on Model Assumptions: Identifying and Conditional Independence

The coefficients from Equation 9 take on a causal interpretation under the assumption that within-families, and conditional on other covariates, the child care decision across siblings is as good as random, and that the treatment effect does not spill over to siblings. The standard test of the identifying assumption is to look for balance in observables across siblings within families. Deming (2009) finds little evidence that Head Start attendance is correlated with observable differences across siblings, which suggests that the magnitude of selection may be small.

In Appendix Table B.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. For the white sample, there are few statistically significant correlations, which suggest that the assumption may be reasonable. For the black sample, participation in Head Start is correlated with a greater likelihood of having higher income at age 1, and lower income at age 2. These correlations 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, which may bias the estimated effects downward. However, given the many hypotheses being tested in this table, it is also possible that these significant findings might be spurious. Moreover, these results are somewhat sensitive, becoming insignificant when we drop observations with imputed controls. We are therefore uncertain how worrying these estimates are.

For the reweighting, we require that CIA holds, which in our context implies that treatment effects should be independent of whether one is in the target population. Appendix Table B.10 examines whether treatment effects vary across individuals in the target population. To implement this test, we first estimate family-specific treatment effects for Head Start. In a second step, continuing to use the switching sample, we regress these estimated treatment effects on an indicator for whether an individual is a member of the target population. These second step regressions are re-weighted for balance on observables.[42] For most outcomes, this test passes, with no sign of systematic differences across target and non-target individuals. However in the CNLSY for the outcomes Learning Disability and Poor Health, and the target population of Head Start participants, there is some evidence ($p < 0.10$) of differential treatment effects for HS participants compared to

[40] Specifically, we use the following individual covariates to generate the propensity score: year of birth, gender, mother's years of education, income at age 3, and income at age 4, and a linear and quadratic in number of siblings. Propensity score estimation is performed separately for the white and black samples. For the post-regression weighting, we then average the ratio of predicted probabilities within families to get family-level weights.

[41] The results between these two methods diverge slightly when individual weights and individual covariates are used. We determined from simulations that in this environment for our data, post-regression weighting typically produces a reweighted estimate closer to the truth, so favor these.

[42] For target individuals the weights are $1/Pr\left[target_i, switcher_i\right]$, and for non-target individuals the weights are $1/\Pr[nontarget_i, switcher_i]$

non-participants. Consequently, we advise that the results for these outcomes and target population be viewed with some caution.[43]

Our reweighting procedure also relies on adequate overlap of the propensity score across switchers and individuals in the target population in the non-switching sample. In Appendix Figure B.3 we show the density of the estimated probabilities of being a Head Start participant for the switching sample and the non-switching Head Start participant sample. This figure shows that there is a good deal of overlap across the two groups, but also that there are a few Head Start participants whose p-scores lie outside the range of the switchers. These observations represent 5 individuals, 6% of the Head Start non-switcher observations, and 3% of all Head Start participants. We interpret this magnitude of violation of the overlap assumption as mild enough to disregard in our subsequent analysis.[44]

6.3 Head Start Results

6.3.1 Reweighted Estimates

We begin by presenting results for our illustrative outcome, attainment of some college for whites in the PSID, in Panel A of Table 5. Column (1) of the table presents the estimated impact of Head Start on some college in GTC, column (2) presents the results using our expanded sample, and columns (3) to (5) present reweighted estimates for the three target populations. As reported earlier, we estimate that Head Start increases the likelihood of attaining some college by a statistically significant 12 p.p. (se: 0.053) using the baseline FFE model. This estimate is 57% smaller than the estimate reported in GTC, 0.281 (se: 0.108).[45] The standard errors are also roughly 50% smaller, corresponding to the roughly tripling of sample size (2,986 compared with 1,036).

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

[43] This test can only be performed among switchers, which requires at least one participant and one non-participant per family. When the target population is Head Start particpants, this requirement forces a degree of blance across the target and non-target groups. Another way of viewing this test is: do switching families with a greater share of participants have different coefficients on Head Start than those with a smaller share of participants? We have run analogous models at the family level, which give qualitatively similar results.

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

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

we should expect the reweighted estimates to show a reduced impact of Head Start relative to FFE.

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

Panel B of Table 5 presents results for the Economic Sufficiency Index in the PSID. Our FFE estimate shows a statistically insignificant 0.023 SD decline in this index associated with Head Start. When we reweight the effects, we find slightly larger negative effects for Head Start eligible children and Head Start participants, and a positive effect (0.03 SD) for siblings. It bears emphasizing, though, that the results are not precisely estimated, such that the 95% confidence intervals allow for a sizeable positive impact of Head Start in spite of the small or negative point estimate. For example, the confidence interval for the economic index for whites allows for a Head-Start-induced improvement of 0.16 SD or a reduction of 0.21 SD for Head Start participants. This limits our ability to make firm conclusions about Head Start's impact on this outcome.

The following four panels of Table 5 show the CNLSY FFE estimates, those reported in Deming (2009) and our replication, and our reweighted estimates. The panels report effects for high school graduation, idleness (not in school or at work), diagnosis of a learning disability, and poor health (based on self-reported health status). The FFE estimates indicate that Head Start leads to a 8.5 p.p. increase in high school graduation ($p < 0.01$), 7.2 p.p. decline in idleness ($p < 0.10$), 5.9 p.p. decline in having a learning disability ($p < 0.01$), and a 6.9 p.p. decline in reporting poor health ($p < 0.01$). The reweighted estimate for participants for high school is 40% smaller, and marginally significant ($p < 0.10$). We also see a substantial 34% decline in the estimated impact on idleness when we consider the impact on participants. The disability and poor health estimates are relatively more stable; the reweighted impacts on participants are just 4% and 25% smaller than the FFE estimate.

In the final column of the table, we test whether the difference between the reweighted estimate for participants and the FFE estimate is statistically significant. We bootstrap the standard errors for this difference by taking draws with replacement from the sample and performing the FFE estimation and reweighting again. We do this 1,000 times and obtain the standard error of our difference as the standard deviation of the 1,000 estimated FFE-reweighted differences. We find that the reweighted estimates for some college (PSID) and high school graduation (CNLSY) are statistically different than the FFE estimate at the 5% and 10% levels, respectively. The remainder of the outcomes are more imprecisely estimated, and therefore we can not reject that the reweighted

estimate is the same as the FFE estimate (in these cases the quantitative differences are non-trivial).

Returning to the PSID, Appendix Tables B.11 and B.12 show the PSID FFE estimates and reweighted results for high school and the good health index for whites, and the corresponding results for blacks. Overall, the results suggest little support for a positive long term effect of Head Start. This is true for the FFE estimates and the reweighted estimates. Nonetheless, the magnitude of the estimates can vary importantly with reweighting, particularly for whites. This makes sense since the identifying sample is a much smaller share of the overall sample for whites relative to blacks. For example, the FFE estimate for the good health index for whites is -0.265 SD, but reweighting for the Head Start participant population changes this estimate to -0.423. In contrast, the coefficients are relatively stable for blacks.[46]

6.3.2 More Evidence on the Role of Family Size

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

We have examined whether the larger coefficients for larger family sizes in Table 3 are driven by family size standing in for other covariates. In Appendix Table B.13 we perform a "horse race" analysis, comparing whether heterogeneous coefficients load on to family size, or other covariates. This table shows that the heterogeneity with family size is robust to also allowing for heterogeity along other covariates. We have also experimented with specifications that test for whether larger family size is merely proxying for "longer sibling cohort span," and do not find evidence that this is the case.

6.3.3 Additional FFE Estimates

Continuing our analysis of the PSID, we also investigate effects of Head Start on a variety of additional short-term outcomes, outcomes at age 40, as well as heterogeneity by race, gender by cohort in Appendix B.1. We do not find any systematic evidence of effects on any of these outcomes, or important heterogeneity along these dimensions.

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

7 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 groups in a panel. In this article, we show that an additional assumption is needed for the *external* validity of results: that panels with variation (switchers) have comparable treatment effects to panels without variation (non-switchers). In other words, fixed effects estimates are generalizable only if there is no *selection into identification*.

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

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

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

Overall, we interpret our findings as pointing primarily toward "increased uncertainty" and to a limited degree toward "zero effects" of the Head Start program. This suggests that there is some discordance between the long-term results from the FFE design, and new estimates using other designs, which generally produce larger and more robust effects of this intervention. We leave it to future research to reconcile these findings.

References

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

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

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

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

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

Andrews, Isaiah, and Emily Oster. 2018. "Weighting for External Validity." National Bureau of Economic Research Working Paper 23826.

Angrist, Joshua, and Ivan Fernandez-Val. 2010. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." National Bureau of Economic Research Working Paper 16566. DOI: 10.3386/w16566.

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

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

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

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91(434): 444–455.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

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

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

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

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

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

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

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.

Sloczynski, Tymon. 2017. "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands." Working Paper.

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

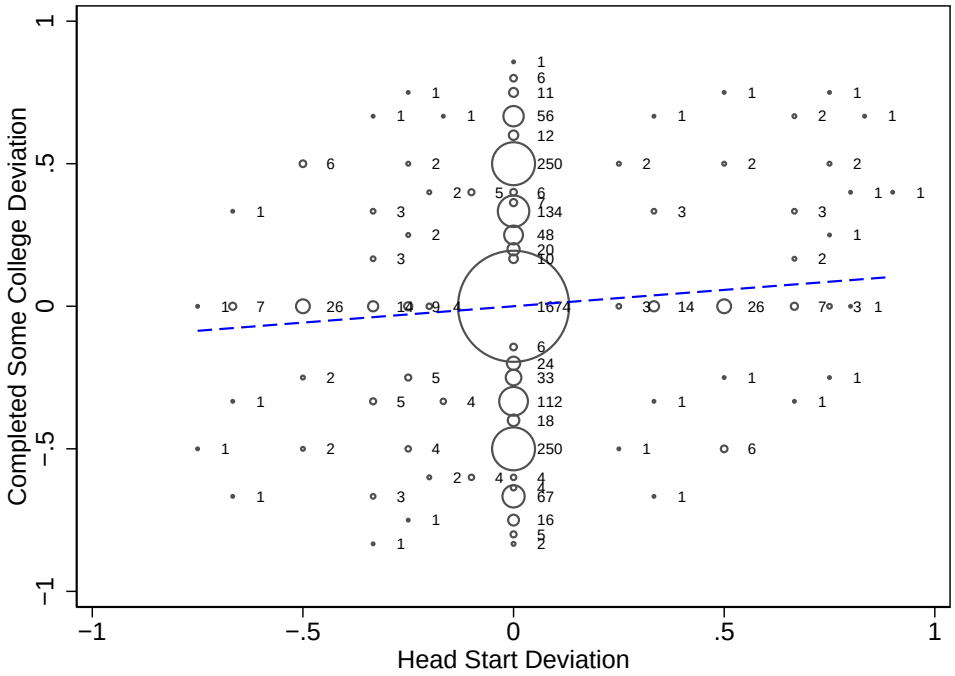
Thompson, Owen. 2017. "Head Start's Long-Run Impact: Evidence from the Program's Introduction." *Journal of Human Resources*.

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

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

8 Figures

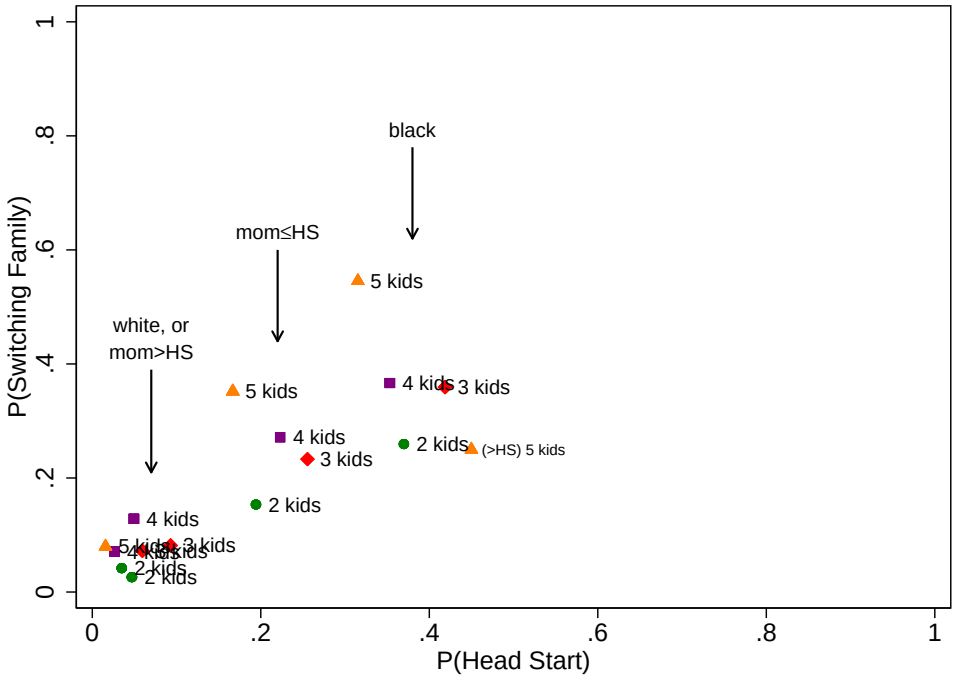
Figure 1: Within-Family Variation in Head Start and Attendance of Some College (PSID)



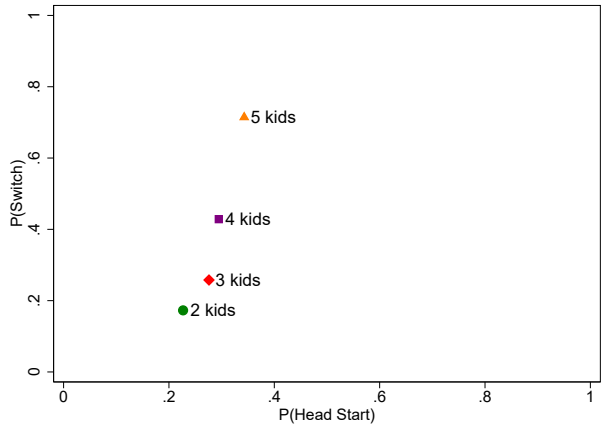
Notes: This figure depicts the identifying variation used in a FFE regression of some college on an indicator for participation in Head Start. Each marker represents the number of individuals that exhibit a particular deviation from the mean Head Start attendance of their family and from the mean attendance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

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

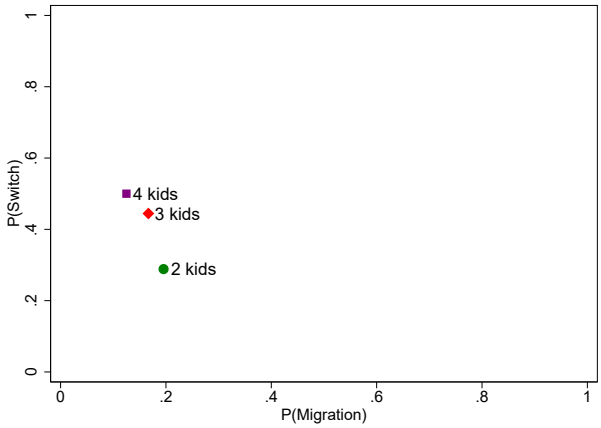
(a) Head Start in PSID



(b) Head Start in CNLSY



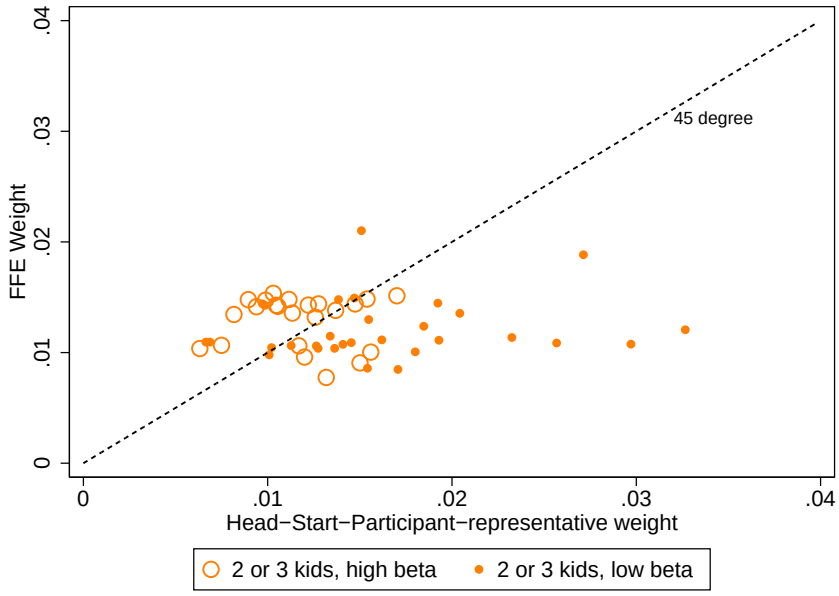
(c) Migration in Census



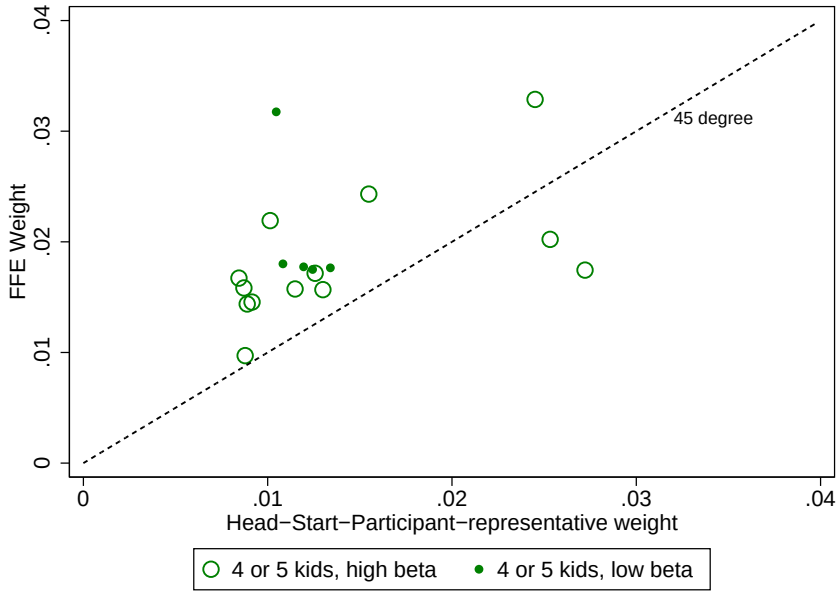
Notes: Panel (a) of this figure plots the *observed* probability of being in a switching family and of attending Head Start by family size for the following groups in the PSID: Whites, Blacks, children of mothers with at most a high school degree, and children of mothers with at least some college. Figure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

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

(a) Families with 2-3 Children



(b) Families with 4+ Children



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

9 Tables

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

	Binary Indep.	Binary Dep.	Both Binary	Total
AEJ: Applied	7	5	4	9
AEJ: Economic Policy	1	1	1	1
AER	3	1	1	5
AER Papers and Proceedings	2	2	1	3
Journal of Health Economics	5	3	2	7
Journal of Human Resources	7	2	2	12
Journal of Labor Economics	2	1	1	5
Journal of Political Economy	3	1	1	3
Journal of Public Economics	4	5	4	6
QJE	1	4	1	4
Review of Economics and Statistics	2	0	0	3
Total	37	25	18	58
Common Dependent Variables				
Schooling/Attainment	24			
Test Score	17			
Employment/Earnings	16			
Birth Weight	8			
Health	6			
Behavioral Issues/Crime	5			
Common Independent Variables				
Schooling	8			
Birth Weight	5			
Health	5			
Parental Traits	4			
Employment	3			
Birth order	3			
Means-Tested Public Program	2			
Death of Family Member	2			
Bombing/Radiation	2			
Observations by Sample				
	Siblings N	Total N		
p10	469	1,212		
p25	1,212	3,255		
p50	6,792	17,501		
p75	175,686	405,802		
p90	750,697	1,582,142		
Year Publication Min/Max	2002	2017		

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. Articles were initially idenitified using the search terms "family," "within family," "sibling," "twin," "mother," "father," "brother," "sister," fixed effect," "fixed-effect," and "birthweight" using queries on journal websites. Siblings N is the number of observations reported for the sample of siblings, while Total N represents the number of total observations reported. See text for details.

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

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

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

Table 3: Returns to Head Start by Family Size,

and Implications for Regression Estimates

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

Notes: Panel A of this table shows the coefficients from a regression of some college on a series of indicators for whether an individual attended Head Start interacted with an indicator for the number of children in one's family. The sample is composed of white individuals. Columns 1 include controls, but not mother f.e., and standard errors are clustered at 1968 family id. Column 2 includes mother fixed effects, and standard errors clustered by mother id. The bottom rows show the weighted average of the coefficients when using regression weights, ω_z (defined in Section 3), determined by the overall distribution of families ("All"), the distribution of 2+ child families ("Siblings"), and the distribution of 2+ child families that have variation in Head Start attendance ("Switchers"). * p < .10, ** p < .05, *** p < .01. Source: Panel Study of Income Dynamics, 1968-2011 waves and Children of the National Longitudinal Study of Youth.

Table 4: Monte Carlo Experiments: Bias of Reweighting and FFE Relative to True ATE, and Efficiency of Reweighting Relative to FFE

		Bias of FE and Reweight:			MSE of Reweight relative to FE:	
	True ATE	FE	Reweight	+ Spline	Reweight	+ Spline
A. Constant TE						
Switchers	80	-0.1	-0.2		1.04	
Siblings	80	-0.1	-0.0		1.16	
All	80	-0.1	-0.2		1.17	
HS Participants	80	-0.2	0.4		1.03	
B. Large Family TE						
Switchers	83.0	-10.6*	0.1		0.93	
Siblings	49.8	22.6*	-0.0		0.68	
All	40.4	32.0*	-0.1		0.52	
HS Participants	41.2	31.3*	-0.1		0.53	
C. TE linear in X_f						
Switchers	93.8	-1.1*	-0.2	0.5	1.04	1.04
Siblings	80.2	12.5*	3.2*	0.6	0.98	1.09
All	80.0	12.7*	3.2*	0.6	0.98	1.10
HS Participants	91.5	1.2*	0.3	0.5	1.03	1.09

Notes: This table shows the results from 10,000 Monte Carlo simulations. Each panel of the table shows results from a different DGP, and each row of the panel is for a different target population. The true DGP is linear, and is discussed in Section 4.4. The first panel shows results where Head Start has a constant treatment effect (TE) for all individuals; the second shows results where Head Start (HS) has no effect on individuals from small families (3 or fewer children) and a large effect for families with many children (4 or more children); and the third panel shows results where treatment effects that are linear in X_f. Column 1, "True Beta," presents the true average increase in the probability of completing some college for participants in Head Start in the sample, which is a function of the DGP and sample composition. Columns 2, 3, and 4 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, and propensity-score reweighted LPM when we include a spline in X_f in the propensity score, respectively. Columns 5 and 6 present the ratio of the mean squared error (MSE) of the two reweighting estimators relative to LPM. Reweighted estimates are obtained using in-regression weighting, with weights adjusting for the representativeness of switchers and the conditional variance of Head Start within families. All betas are multiplied by 1,000. * $p < .01$.

Table 5: Head Start Impact for Representative Eligible Children, Participants, and Siblings

Using Post-Regression Reweighting Method

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

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

A Appendix: Derivation of Propensity Score Weighting

A.1 Derivation of Propensity Score Weighting

In this section, we provide a simple derivation of the weighting scheme that we propose to obtain the ATE from the switchers sample. We introduce a concrete example in which treatment effect is determined by one variable, race, and in which there are only few groups in the switcher sample, and focus on families as the group-level variation of interest. This makes it easy to derive the share of the target population corresponding to each individual in the switcher sample.

A.1.1 Thought Experiment

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

A.1.2 Notation

Under the setup above, the weight that should be given to a switcher family f with race r, s_{fr}, can be written as:

$s_{gr} = \frac{N_r}{N_{target}} \times \frac{1}{N_{r,switch}} \times n_g$

where N_r is the number of individuals in the target population with race r, N_{target} is the number of individuals in the target population, $N_{r,switch}$ is the number of individuals with race r in the switcher sample, and n_g is the number of individuals in family g.

This is equivalent to:

$s_{gr} = pr(race|target) \times \frac{1}{pr(race|switch) \times N_{switch}} \times n_g$

$s_{gr} = \frac{pr(race|target)}{pr(race|switch)} \times \frac{n_g}{N_{switch}}$

$s_{gr} = \frac{pr(race|target)}{pr(race|switch)} \times pr(g|switch)$

A.1.3 Estimation

We can obtain an estimates of this as follows:

1. We obtain an estimate of $pr(target|race)$ as fitted values from a regression of being in the target on b.

   This is equal to $\frac{pr(race|target) * pr(target)}{pr(race)}$ by Bayes rule.

2. We obtain an estimate of $pr(switch|race)$ as fitted values from a regression of being a switcher on b.

   This is equal to $\frac{pr(race|switch) * pr(switch)}{pr(race)}$ by Bayes rule.

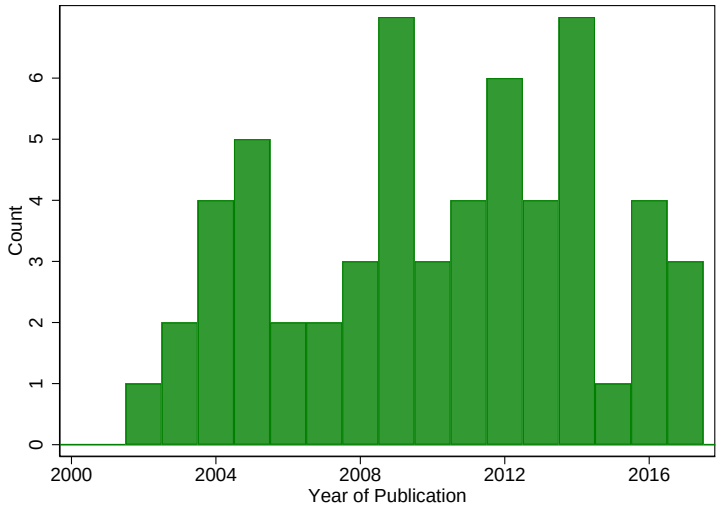
Taking the ratio of these, we have $\frac{pr(race|target)}{pr(race|switch)} \times \frac{pr(target)}{pr(switch)}$.

To obtain s_{fr} we need to multiply this ratio by $pr(g|switch)$, $\frac{n_g}{N_{switch}}$, and divide by $\frac{pr(target)}{pr(switch)}$. Since $\frac{pr(target)}{pr(switch)}$ is constant across all families, we can operationalize this through normalization of the weights.

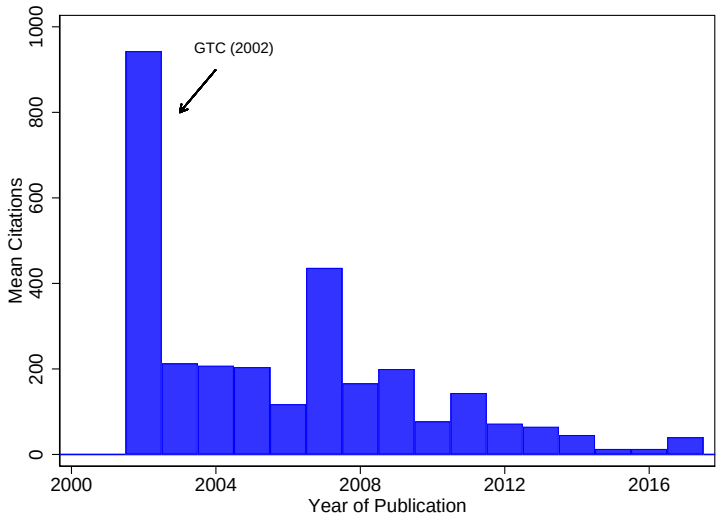
B Appendix: Supplementary Figures and Tables

Figure B.1: Popularity of Family Fixed Effects Articles

(a) Publications by Year



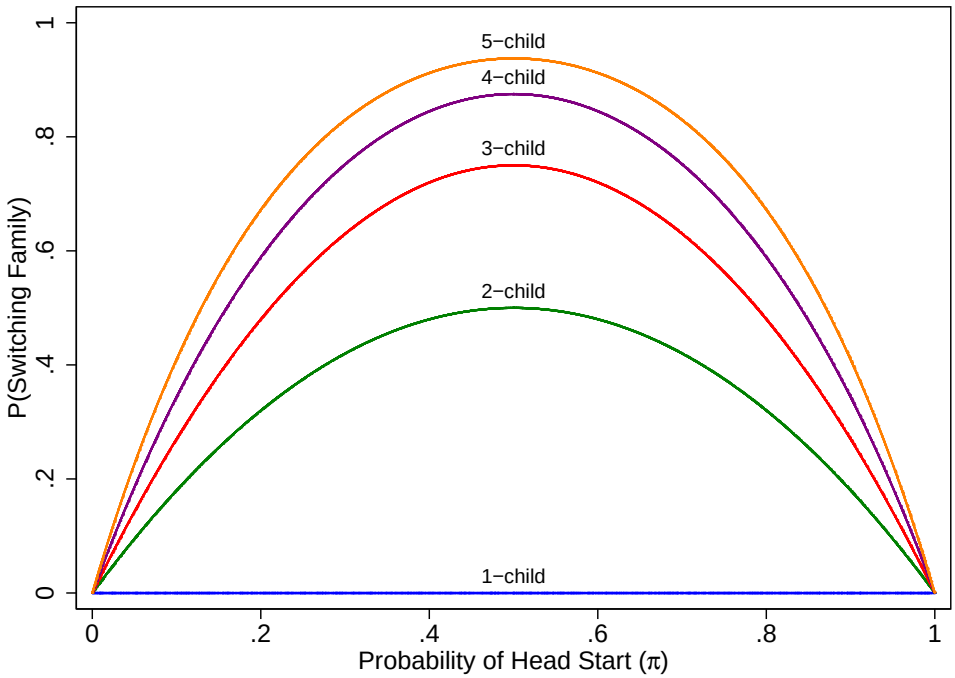
(b) Average Citations by Year of Publication



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

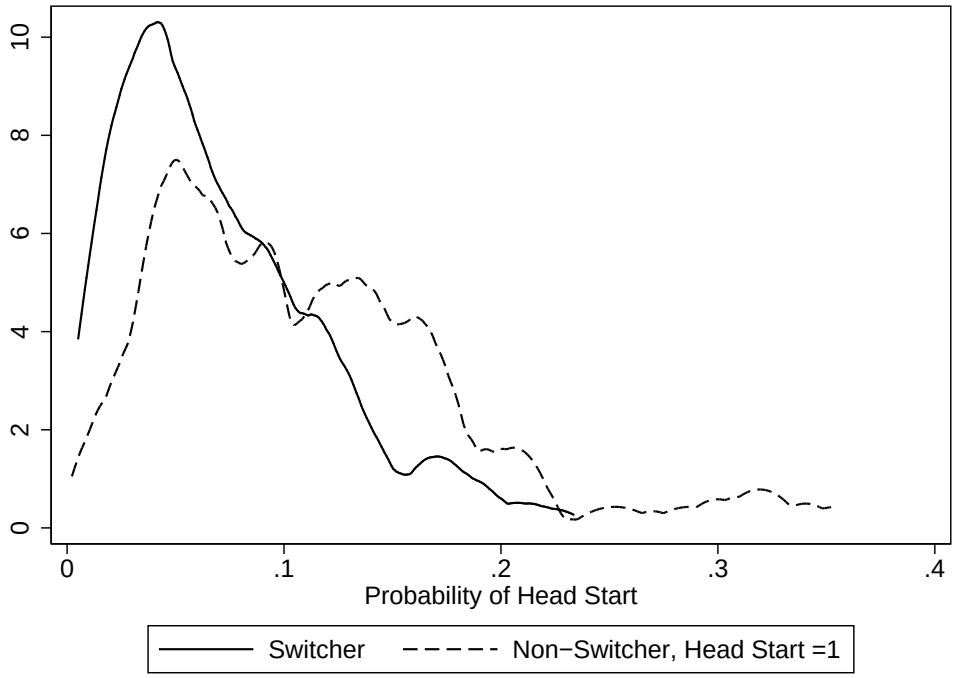
Figure B.2: Illustrative Model of the Role of Family Size in Switching

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



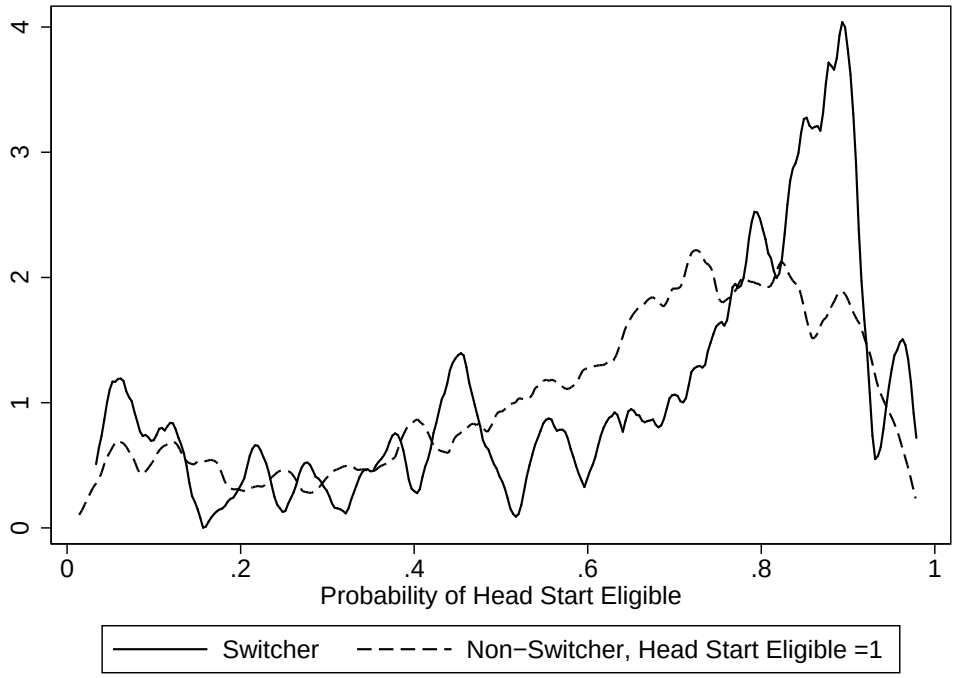
Notes: This figure plots the theoretical function: $P(HSSwitchingFamily) = 1 - (1-\pi)^{n_g} - \pi^{n_g}$, where n_g is the number of children in a family and π is the probability of attending Head Start, for 2-, 3-, 4-, and 5 (plus)- child families.

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



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

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



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

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

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

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

Table B.2: Change in Weighting of Regression Estimates Across Sibling and Switcher Samples (PSID)

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

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

Table B.3: Demographic Characteristics of Head Start Sample (PSID)

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

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

Table B.4: Outcomes of Interest for Head Start Sample (PSID)

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

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

Table B.5: Summary Statistics for Inputs to Summary Indices (PSID)

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

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

Table B.6: N's for Control Covariates (PSID)

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

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

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

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

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

Table B.8: N's for Auxiliary Outcomes (PSID)

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

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

Table B.9: Effect of Head Start on Pre-Head-Start Outcomes (PSID)

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

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

Table B.10: Test of Conditional Independence Assumption:

Do Individuals in the Target Population Have Differential Treatment Effects?

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

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

Table B.11: Additional Outcomes for Representative White Eligible, Participants, and Siblings (PSID)

Using Post-Regression Reweighting Method

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

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

Table B.12: Head Start Impact for Representative Black Eligible, Participants, and Siblings (PSID)

Using Post-Regression Reweighting Method

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

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

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

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

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

B.1 Supplementary PSID FFE Results

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

We present the FFE results for the economic and health indices measured at age 40, together with the indices at age 30 for comparison, in Table B.14. Overall, the results suggest little support for a positive long term effect of Head Start. We come to the same conclusions when we aggregate the inputs using principal components analysis (see Table B.15. Our overall conclusions are not changed importantly by looking at specific outcomes or subsamples. We have also estimated regressions for each of the inputs to the economic and health indices, which we include in Tables B.16, and B.17, B.18, B.19. Table B.20 shows the regression results for the additional outcomes analyzed in GTC, earnings between ages 23 to 25, and not having committed a crime. Across these tables, there is no systematic evidence that Head Start impacts long term outcomes.[47]

Motivated by the prior findings of differential effects by gender in Carneiro and Ginja (2014); Deming (2009), in Table B.21 we look to see whether our mean results are obscuring this form of heterogeneity in our setting. Curiously, we find some evidence of significant negative effects of Head Start among men, in particular for health and economic outcomes at age 40. On the other hand, we find a positive and significant effect of Head Start on the probability that men attain some college. The effects estimated for women are never individually significant, but also not statistically different from men for many outcomes as indicated by the p-value of the difference in the table. The one exception is for economic outcomes observed at age 40, where women are found to have significantly better returns to Head Start participation than observed for men.

Another source of heterogeneity which could generate a discrepancy between our results and GTC is the fact that our sample includes later (younger) cohorts, whose Head Start experience may differ from earlier participants. In Table B.22, we find some support for a decreasing impact of Head Start across cohorts for the age 40 indices, but also find a larger improvement in the health index at age 30 for more recent cohorts. Thus, this does not appear to reconcile our findings.[48]

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

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

Table B.14: Impact of Head Start on Economic Sufficiency Index and Good Health Index (PSID)

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

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

Table B.15: Effect of Head Start on Economic and Health Principal Components (PSID)

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

Notes:

Table B.16: Effect of Head Start on Inputs to Economic Sufficiency Index at age 30 (PSID)

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

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

Table B.17: Effect of Head Start on Inputs to Economic Sufficiency Index at age 40 (PSID)

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

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

Table B.18: Effect of Head Start on Inputs to Good Health Index at age 30 (PSID)

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

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

Table B.19: Effect of Head Start on Inputs to Good Health Index at age 40 (PSID)

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

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

Table B.20: Impact of Head Start on High School, College, Earnings, and Criminal Behavior (PSID)

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

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

Table B.21: Impact of Head Start on Main Outcomes by Sex

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

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

Table B.22: Regression: Interaction with Cohort (Linear) (PSID)

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

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

C Appendix: Replication of GTC (2002)

C.1 Summary

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

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

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

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

C.2 Replication Methodology

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

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

C.2.1 Construction of Dataset

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

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

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

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

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

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

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

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

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

C.2.2 Search for identical dataset construction

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

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

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

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

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

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

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

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

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

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

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

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

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

C.3 Final dataset restrictions

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

C.4 More on the data

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

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

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

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

Table C.2: GTC Table 1: Summary Statistics

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

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

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

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

Table C.4: GTC Table 2: Regressions

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

SE in parentheses.

Table C.5: Alternative Definitions of Race

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

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

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

Table C.7: Iterations for Summary Statistics Table

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

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

Table C.8: Iterations for Regressions Table

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

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

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

Table C.9: PSID Variables used in the analysis

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

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

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

D Functional form choices with Binary Treatment and Binary Outcome

We now consider potential sensitivity to functional form modeling assumptions. For binary outcomes the usual choice of specifications include linear probability model (LPM), logit, and probit. In the cross-sectional setting, the conventional wisdom is that the choice among these options is fairly innocuous, especially when the objective is to recover the ATE.[49] We are not aware of any previous systematic exploration of these properties in extremely short-panel settings such as found in the FFE design. We demonstrate some complications that arise in such settings, and compare the performance of these estimators.

D.1 Specification choices

Empiricists commonly use LPM specification to estimate FE models. In our sample of papers, this is almost universally used as the primary, if not only, specification. We speculate that this is motivated by (1) the intuition carried over from the cross-sectional case that LPM models usually recover the ATE; (2) the benefit that the incidental parameters problem does not pollute the main parameters of interest (Chamberlain, 1980);[50] (3) computational ease, especially when paired with other complications to the research design such as many fixed effects, instrumental variables, etc.); and (4) the fact that the estimated coefficient β_{LPM} directly gives the estimate of the ATE.

Obtaining ATE from a nonlinear specification is not only less common, but also sometimes less straightforward. The conditional logit model, sometimes referred to as logit FE, consistently estimates β_{Logit} by conditioning on the number of successes in a family, but does not have a paired method for obtaining treatment effects. To obtain ATE, Wooldridge (2010, section 15.8) recommends employing a regular logit model and including family-level-means of control variables, i.e. "Chamberlain-Mundlak controls," (hereafter, Mundlak controls) rather than directly controlling for fixed effects (Mundlak, 1978; Chamberlain, 1980).[51]

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

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

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

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

we face extremely short panel lengths due to families commonly having only 2 children[52]; (2) his results apply only to the leading (order 1/T) bias term, but with very short panels the subsequent bias terms could still be relevant; and (3) there is an unresolved challenge of how to address the extrapolation from the estimation sample to singletons (when they are in the target population).[53]

Mundlak controls and naive-fixed-effects methods have the attractive properties of: (1) being easy to implement; (2) respecting the binary functional form of the left-hand-side (LHS) variable; and (3) straightforwardly obtaining ATEs. Nonetheless, empiricists' use of either of these options is uncommon; in our sample of 58 papers discussed in section 2.2 these methods are not used.

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

D.2 Obtaining Marginal Effects from Conditional Logit

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

$$ATE_{2StepLogit} = \frac{1}{N}\sum_{i=1}^{N}\int_{u}(\hat{\beta}_{HeadStart}\cdot\Lambda(\hat{\beta}X_{if}+\hat{\gamma}Z_f+u)\cdot 1-\Lambda(\hat{\beta}X_{if}+\hat{\gamma}Z_f+u)\phi(u)du \quad (10)$$

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

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

With $\hat{\beta}_{HeadStart}$ the coefficient on Head Start from the conditional logit first step; $\hat{\beta}$ the coefficient on i-level variables X_{if} from the conditional logit first step; $\hat{\gamma}$ the coefficients on family-level variables from the second step; and $\phi(u)$ the PDF from a normal distribution, with variance σ_u^2 estimated from the second step family-level random effects model. We have not yet found a prior implementation of this estimator in the literature; but it is similar in spirit to the two step fixed-effects logit proposed by Beck (2015).[54]

D.3 Selection in Nonlinear Models

A desirable feature of the two-step logit and the Mundlak controls models is that both allow the marginal effect of treatment, $\frac{\partial Pr(Y=1)}{\partial HeadStart}$, to vary across individuals. In both cases, the treatment effect depends on the "index value" for each individual. However, the models maintain an assumption of constant treatment effects in logit units (β_{Logit}). If the model is misspecified, and instead there are variable treatment effects for different individuals, and that a reweighted estimation sample might produce more reliable results, especially when trying to measure the ATE for a pre-specified target group. This consideration is analogous to the treatment effect heterogeneity discussed in section 4.3.

We propose employing the in-regression weights $\widetilde{s_f^{sw \to tg}} \cdot v_f$ as discussed above in section 4.3. That is, the weights are a combination of (1) propensity score weights derived from a multinomial logit model predicting "RHS switcher" status and "in target population" status, and (2) inverse within-family conditional variance of the treatement variable of interest. For expediency, we continue to estimate this conditional variance from a linear model, and to apply it directly to the second stage logit estimation step.

We explore some of these models in the context of our empirical example, and find some differences in the point estimates and precision across linear and nonlinear specifications. Compared with LPM, we find somewhat smaller and less precise impacts of Head Start on some college when we use the 2-step approach (point estimate: 0.086 (se: 0.059). We note that the slight decrease in precision here accompanies many fewer observations, which has fallen to 1200 for estimation of the logit beta instead of 2987 in the LPM.[55] The point estimate for the Mundlak controls is very similar to LPM, 0.126, but the standard errors are 20% larger (se: 0.053), so that the estimate is significant only at the 10% level.

D.4 Monte Carlo for Nonlinear Specifications

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

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

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

For our simulations, we continue with the PSID data setup presented above in Section 4.4.1. We take the original data, and estimate a logit model predicting some college attainment, using as regressors family level variables and family-level averages of individual variables. From this model we construct a family level logit index variable, x_f. For each simulation, the underlying logit index for each individual is equal to x_f, plus the Head Start dummy multiplied by the Head Start (logit) treatment effect. We then turn this index into $Pr(y = 1)$ using the logistic CDF, and then randomly draw outcomes y. We consider three DGPs. The first of these has a constant treatment effect (in logit units). The second has a treatment effect that is zero for small families (with 2 or 3 children), and a larger treatment effect for families with 4 or more children. The third DGP has a variable treatment effect which is decreasing linearly in x_f. For all of the DGPs, the treatment effect in terms of $Pr\,(y = 1)$ will vary across target populations because different children have different logit indices. For DGPs 2 and 3, there is additional variability stemming from family characteristics.

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

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

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

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

D.5 Discussion of Specification Choices

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

estimation first step) or a logit with Mundlak controls can perform well.

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

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

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