Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care-Type Settings*

Avi Feller   Todd Grindal   Luke Miratrix   Lindsay Page

Initial Draft: September 1, 2014
This Draft: January 11, 2016

Abstract

Early childhood education research often compares a group of children who receive the intervention of interest to a group of children who receive care in a range of different care settings. In this paper, we estimate differential impacts of an early childhood intervention by alternative care setting, using data from the Head Start Impact Study, a large-scale randomized evaluation. To do so, we utilize a Bayesian principal stratification framework to estimate separate impacts for two types of Compliers: those children who would otherwise be in other center-based care when assigned to control and those who would otherwise be in home-based care. We find strong, positive short-term effects of Head Start on receptive vocabulary for those Compliers who would otherwise be in home-based care. By contrast, we find no meaningful impact of Head Start on vocabulary for those Compliers who would otherwise be in other center-based care. Our findings suggest that alternative care type is a potentially important source of variation in early childhood education interventions.

1 Introduction

Access to publicly funded prekindergarten in the United States has expanded substantially in recent years. In the last decade, the percentage of U.S. four-year-old children enrolled in public preschool has increased by one-third—from 31 to 40 percent—with some states now serving nearly 90 percent of all four-year-old children through publicly funded preschool programs (Barnett et al., 2014). Many cities, such as Boston, Los Angeles, New York, and Washington, D.C., have added to this expansion through locally-funded prekindergarten programs. The Obama Administration has called for additional funds to support even greater access to high-quality early childhood education across the country.

Those who support the expansion of publicly funded preschool point to nearly 50 years of research indicating that participation in high-quality pre-school programs can yield individual and

[*] *Email:* afeller@berkeley.edu. We gratefully acknowledge financial support from cooperative agreement #90YR0049/02 with the Agency for Children and Families (ACF) of the U.S. Department of Health and Human Services and from the Spencer Foundation through a grant entitled "Using Emerging Methods with Existing Data from Multi-site Trials to Learn About and From Variation in Educational Program Effects." All opinions expressed in the paper and any errors that it might contain are solely the responsibility of the authors.

societal benefits in both the short and long term, often highlighting historically important interventions such as the Perry Preschool Project (e.g., Barnett, 1995; Heckman, 2006). Opponents argue that current public preschool programs, especially Head Start, the largest and most prominent public preschool program in the United States, have failed to replicate these initial successes at scale (e.g., Coulson, 2013; Whitehurst, 2013b). This belief stems in part from the results of the Head Start Impact Study (HSIS), a randomized evaluation that found that the opportunity to enroll in Head Start improved children's performance on short term measures of cognitive and social-emotional development but that, in general, these initial impacts were no longer apparent after children finished first grade (Puma et al., 2010a).

Researchers and policymakers have posited a wide range of explanations for differences between the Head Start results and those of early model programs like Perry preschool, including differences in program features, program intensity, and program targeting (Barnett, 2011; Bitler et al., 2014; Elango et al., 2015). We focus on one prominent explanation: that the care settings of control group children attenuated the reported effects for Head Start (e.g., National Forum on Early Childhood Policy and Programs, 2010). In the Perry Preschool Project, all control group children were cared for in their homes by a parent or other adult. By contrast, in the Head Start Impact Study, roughly one-third of children not in Head Start enrolled in other center-based care, with services similar to those provided by Head Start, instead of receiving care in a home-based setting.

In this paper, we conduct a comprehensive analysis of the differential impact of enrolling in Head Start by the setting in which children would otherwise receive care. Our main result is that enrollment in Head Start yields strong, positive short-term effects on a measure of receptive vocabulary among those children who would enroll in Head Start when offered the opportunity to do so but who would otherwise be cared for by a parent or other caregiver at home or in a home-based setting. For this group of children, we estimate that, after one year, enrollment in Head Start improved children's performance by over 0.2 standard deviations, more than 50 percent larger than the corresponding intent-to-treat estimates reported in Puma et al. (2010a). By contrast, we find no meaningful impact of Head Start for those children who would otherwise enroll in non-Head Start center-based care.[1]

Our analysis makes three main substantive contributions. First, we find meaningful impact variation by alternative care type that is masked by the HSIS topline results. This suggests that sweeping claims of Head Start's ineffectiveness (e.g., Whitehurst, 2013a) are misplaced, at least in terms of impact on receptive vocabulary. At the same time, we find no evidence that other center-based alternatives are more effective than Head Start on average, despite research arguing that this might be the case (Gormley et al., 2010). Second, this pattern of impact variation broadly holds across outcome quantiles (Bitler et al., 2014) and within key subgroups (Bloom and Weiland, 2014), although these estimates are imprecise. We find especially large impacts among Dual-Language Learner children who would otherwise be in home-based care. Third, consistent with the HSIS

[1] These results are corroborated in independent work by Kline and Walters (2015), who find the same general pattern using a structural model. We compare our approaches in Section 7.

results (Puma et al., 2010a), we find that, while the impact of Head Start indeed declines over time, it is a gradual decline rather than the rapid attenuation identified by prior work (Gibbs et al., 2011). We also find modest evidence of positive impacts of Head Start through first grade.

Our paper also makes several methodological contributions. First, we set up an approach for identifying and estimating impacts in the presence of multiple counterfactual treatment options, which is common in early childhood education studies and in program evaluation more generally (e.g., Heckman et al., 2000; Duncan and Magnuson, 2013). To do so, we use the *principal stratification* framework of Frangakis and Rubin (2002), which is a generalization of the usual instrumental variables (IV) approach for non-compliance in randomized experiments (Angrist et al., 1996). In the standard IV case, the goal is to estimate the impact of randomization for Compliers, known as the Local Average Treatment Effect (LATE). In HSIS, Compliers are children who would enroll in Head Start under treatment and would not enroll in Head Start under control. In our analysis, we are instead interested in two different types of Compliers: Center-based Compliers, children who would enroll in Head Start under treatment and would enroll in other center-based care under control, and Home-based Compliers, children who would enroll in Head Start under treatment and would otherwise enroll in home-based care. This approach yields two LATEs, rather than just one.

Identifying and estimating impacts for these subgroups is challenging. Extending results from the IV setting (Imbens and Rubin, 1997b; Abadie, 2003), we first show that a range of quantities of interest can be immediately estimated using moment-based methods, including the relative sample shares of Center- and Home-based Compliers and the outcome distributions for these groups under control. The outcome distributions under treatment, however, are more difficult to estimate. To overcome these obstacles, we therefore utilize a hierarchical Bayesian modeling approach (e.g., Imbens and Rubin, 1997a). In addition to providing a natural paradigm for causal inference with potential outcomes, this approach easily allows us to account for many of the real-world complications in the Head Start Impact Study, including missing data and a multilevel structure, with children nested within Head Start centers. We estimate this model via an implementation of Hamiltonian Monte Carlo called Stan (Stan Development Team, 2014), which builds on recent advances in Bayesian computation. To the best of our knowledge, this is the first implementation of a principal stratification model with site-level random effects.

We organize the paper as follows. Section 2 gives background on Head Start and the principal stratification approach. Section 3 describes the HSIS data. Sections 4 and 5 provide an overview of the analytic framework and give some descriptive information about the principal strata. Section 6 gives an overview of our identification and estimation approaches. Section 7 presents our results. We close with a discussion of the substantive implications for this work for early childhood policy and reflect on the broader methodological implications. We defer all detailed technical discussions and proofs to the appendix.

2 Background

2.1 Background on Head Start and the Head Start Impact Study

Originally launched in the summer of 1965 as a two-month intervention to help low-income children prepare for kindergarten, Head Start programs across the United States currently provide early childhood education and family support services to more than 900,000 low-income children and their families each year. Head Start services are administered by nearly 1,600 local grantee agencies that receive a total of $8 billion in annual state and federal funds (Administration for Children and Families, 2014). Today, Head Start programs must adhere to a set of performance standards that specify requirements for program services, curricula, teacher preparation and professional development. For example, current Head Start classes serving four or five-year-olds can have no more than 20 children, and those serving three-year olds can have no more than 17 children. Programs must screen all enrolled children for developmental, sensory, and behavioral disabilities and have a written curricula to support each child's cognitive and language development. Head Start programs are also required to engage in collaborative partnership-building with parents through processes that include structured home visits, parenting education classes, and assistance in accessing food, housing, clothing, and transportation.

Researchers and policy makers have debated the effectiveness of Head Start since the program's inception. In their summary of the initial research on Head Start from the 1960s, Zigler and Muenchow (1992) show that children enrolled in these early evaluations of Head Start exhibited large gains on measures of cognitive achievement between their initial enrollment and program completion. Excitement regarding these impressive findings was soon tempered, however, by additional research indicating that the effects of Head Start participation were no longer apparent once children reached elementary school (Westinghouse Learning Corporation, 1969). Nevertheless, many of the quasi-experimental studies that followed over the next four decades indicated positive impacts of Head Start on a range of outcomes from short-term academic skill development to long-term outcomes measured in adulthood (e.g., Currie and Thomas, 1993; Garces et al., 2002; Ludwig and Miller, 2007; Deming, 2009; Carneiro and Ginja, 2014).

The mixed results of the randomized Head Start Impact Study did little to settle this debate (National Forum on Early Childhood Policy and Programs, 2010). Nonetheless, the rich HSIS data has led to a host of secondary analyses. Bloom and Weiland (2014) and Walters (2015), for example, examine impact variation across Head Start centers, finding substantial heterogeneity. Bitler et al. (2014) use quantile regression to examine impact variation across the entire outcome distribution, finding substantially larger effects for children with low scores. Bitler et al. (2014) and Bloom and Weiland (2014) also examine heterogeneity across important subgroups, with both studies highlighting significantly larger effects among Dual-Language Learners than among native English speaking students. Finally, other studies, such as Gelber and Isen (2013) and Miller et al. (2014), find that parents play an important role in the effects of Head Start.

2.2 Heterogeneity by Alternative Care Type

The goal of this paper is to explore a specific type of impact heterogeneity: whether or not the impact of Head Start varies by alternative care type. There is substantial evidence in the literature suggesting that this might be the case. First, a recent meta-analysis of 28 studies of Head Start conducted between the program's inception and 2007 found that much of the variation in the findings regarding Head Start's impact on child achievement and cognitive development could be explained by differences in the types of preschool services used by the control group (Shager et al., 2013). Although studies of Head Start programs yielded overall positive effects on short term indicators of children's cognitive skills and achievement, with average effect sizes of +0.27, those studies in which the children in the control group experienced other forms of center-based care yielded significantly smaller effects as compared to those studies of Head Start in which control group children received no additional services (see also Duncan and Magnuson, 2013, for a broader discussion of the counterfactual problem). Zhai et al. (2011) find a similar result using longitudinal data from the Fragile Families and Child Wellbeing Study, concluding that impacts of Head Start were largest relative to non-center-based care.

Second, a few authors have used HSIS data to address this question. Using a matching approach, Zhai et al. (2014) find significant effects of Head Start compared to parent care and relative/nonrelative care but find no meaningful differences in outcomes between Head Start and other center-based care. Using variation across sites, Walters (2015) finds that impacts are smaller for Head Start centers that draw more children from other center-based programs rather than from home-based care. Finally, using a structural model, Kline and Walters (2015) find that the effects of Head Start are larger relative to home-based care than relative to other center-based care. We discuss the relationship between our results and those of Kline and Walters (2015) in Section 7.

At the same time, some authors have argued against alternative care type as an important source of impact variation. Bitler et al. (2014), for example, find no relationship between observed impacts and the distribution of counterfactual care type across a range of subgroups in HSIS. Barnett (2011) points to the Abecedarian study, initially launched in 1972, which demonstrated large, sustained program impacts, even though roughly two-thirds of control group children attended high-quality center care.

2.3 Principal Stratification

There is a small but growing literature on the use of model-based principal stratification in social science applications. Page et al. (2015) provide a recent non-technical review (see also Schochet et al., 2014). Some previous education examples include Barnard et al. (2003) on the effect a randomized lottery for private school voucher use in New York City with complex noncompliance patterns (see also, Jin and Rubin, 2009); Page (2012) on the relative importance of student exposure to the labor market in career academy high schools; and Schochet (2013) on student mobility in school-based randomized trials. Outside of education, several studies have used principal stratification to analyze the JobCorps evaluation (e.g., Zhang et al., 2009; Frumento et al., 2012) and JOBS II evaluation (Mattei et al., 2013). Finally, a separate series of papers use a *principal score* approach, rather than model-based inference, to estimate similar quantities of interest. The key assumption with this approach is *principal ignorability*: conditional on covariates, stratum membership is ignorable. Examples include Hill et al. (2002), who analyze the Infant Home Development Program, Schochet and Burghardt (2007), who analyze the JobCorps data, Jo and Stuart (2009), who analyze the JOBS II data, and Scott-Clayton and Minaya (2014), who analyze student employment data.

3 Head Start Impact Study

3.1 Overview

Our primary source of data is the HSIS, which was conducted within oversubscribed Head Start centers throughout the U.S. In the HSIS, children randomized to treatment were offered enrollment in a Head Start program for the 2002-2003 school year, while children randomized to control were not offered enrollment. In total, 4,440 children, aged either three or four years old, were randomized to treatment or control across 351 Head Start centers. We exclude all children from Puerto Rico, because they are not available in the public use data set. The randomization itself was complex; treatment probabilities varied by the child's age, the date the child was first put on a Head Start center wait list, and the distribution of eligible children across neighboring Head Start centers.[2] While it is infeasible to recreate the true randomization procedure using currently available data, we can approximately account for the complex structure of the randomization by analyzing the data as if randomization were conducted separately within each center. After excluding children from centers that did not have at least one child in each experimental condition, we obtain a data set with 4,385 children across 340 Head Start centers. We refer to the first year of the study as the Head Start year.

3.2 Outcomes

The HSIS research team collected a wide array of outcomes on children in the sample. A key requirement of our analytic approach, however, is the ability to find a close parametric approximation to the underlying outcome distribution. Therefore, we currently cannot assess several important cognitive outcomes, such as the Woodcock-Johnson III Applied Problems test, and social-emotional outcomes, such as externalizing behavior, since they are poorly suited to typical parametric approximations, even conditional on covariates.

We therefore restrict our analysis to the Peabody Picture Vocabulary Test (PPVT), a standardized measure of children's receptive vocabulary in which the evaluator shows the child a page containing three to four pictures and asks the child to identify the picture that best represents

[2] The official HSIS report also uses a complex set of weights to extrapolate the experimental results to a "nationally representative" population of potentially eligible Head Start children (see Gibbs et al., 2011, for a discussion). We do not use those weights here, instead focusing on the results for the experimental sample.

the meaning of a word presented orally by the assessor. See Puma et al. (2010b, section 3-10) for additional details on the exact form of the PPVT. The PPVT is our variable of choice for two reasons. First, the PPVT, which is derived from an item response theory score, is unimodal and roughly bell-shaped. Second, the PPVT is a widely used assessment and is predictive of key skills later in life (Romano et al., 2010). Based on results from the pre-test, the average child at the beginning of the HSIS performed at roughly the 30th percentile of national PPVT performance, reflecting this group's relative disadvantage in pre-academic skills.

An important complication in the HSIS is the high proportion of missing outcomes. Overall, around 18 percent of PPVT scores are missing in the Head Start year, increasing to around 22 percent two years later. Twenty five percent of PPVT pre-test scores are missing. Furthermore, treatment group children are much more likely to have observed outcomes than control group children: in the Head Start year, 24 percent of control group children have missing PPVT scores, compared to just 13 percent of treatment group children. Around 40 percent of children are missing at least one PPVT score from the pre-test, the Head Start year, the first follow-up year, or the second follow-up year; around 10 percent do not have an observed PPVT score for any of these four tests.

3.3 Covariates

Covariates play a particularly important role in principal stratification models. Zhang et al. (2009) point to two main functions. First, covariates can be predictive of the outcome and stratum membership. Second, parametric assumptions can often be more plausible conditional on covariates than marginally. For additional discussion, see Hirano et al. (2000); Jo (2002); Jo and Stuart (2009); Ding et al. (2011); Feller (2015).

Thankfully, the HSIS data set includes a rich set of covariates on child and family characteristics. As part of a broader research effort, we also appended center-level characteristics and neighborhood-level variables for the area around each child's Head Start center of random assignment. Neighborhood-level information includes geocoded data from the 2000 Census, the 2002 Business Census, the Department of Education, and the FBI crime database (McCoy et al., 2014).

Table 1 assesses balance across conditions for the HSIS covariates we use in our analysis. The left column shows the covariate mean for those children assigned to the control group. The middle column shows the difference between covariate means in the treatment and control groups. Finally, the right column shows the normalized differences, a standardized measure of covariate balance across treatment conditions (Imai et al., 2008; Imbens and Rubin, 2015). There is excellent covariate balance between treatment and control groups, with all normalized differences below 0.1 in absolute value.

Overall, HSIS children had diverse background characteristics (reporting control group means for simplicity): around 30 percent identified as Black, 37 percent as Hispanic, 29 percent spoke a non-English language at home, roughly half lived with both biological parents, and one-fifth had a mother who was a recent immigrant. The children generally come from disadvantaged households: around 70 percent have a mother with at most a high school degree or GED, and around 80 percent have an assessed family risk that is moderate to high.[3] As would be expected, the children's households are generally situated in disadvantaged neighborhoods. Based on the census data for the Head Start centers, nearly one-quarter of neighborhood households were in poverty. Further, while the national unemployment rate in the US was roughly four percent in 2000, the unemployment rate in these communities was nearly eleven percent, although there is substantial heterogeneity across neighborhoods (McCoy et al., 2014).

3.4 Child Care Setting

Standard practice in early childhood education research is to divide care settings into home-based versus center-based care (e.g., Gormley, 2007). Given our main substantive question, we therefore categorize care settings into three main groups: Head Start, non-Head Start center care, and home care. Home care encompasses a variety of home-based settings including being cared for by a parent at home (73 percent), being cared for in a non-relative home-based child care setting (11 percent), being cared for by a relative in that relative's home (9 percent), and being cared for by a non-parent in the family's home (6 percent). Although it may be of some substantive interest to separate out these different home-based settings, it was not feasible given the small sample sizes.

Table 2 shows the distribution of observed child care settings in the Head Start year for children in the HSIS treatment and control groups. Among treatment group children, 77 percent took up the offered slot and enrolled in Head Start in the treatment year. Approximately eight percent of children assigned to treatment enrolled in a non-Head Start center, and nine percent were cared for by a parent or other relative or enrolled in a home-based childcare program. In principle, children randomized to the control group were free to take up any available early childhood program except for that provided by the Head Start center to which they had applied and had not been offered enrollment. In practice, among control group children with an observed care setting, 13 percent enrolled in a Head Start center (most in the center in which they had lost the lottery), 31 percent enrolled in a non-Head Start center, and 56 percent were cared for by a parent, a relative, or within a home-based childcare program. Note that the HSIS sample consists entirely of families who actively sought to enroll a child in Head Start. Thus, there was at least some initial indication of a preference for Head Start.

4 Analytic Framework

We next outline the technical aspects of our Bayesian principal stratification framework. We begin with a general setup for the problem, review the case with binary treatment compliance—that is, Head Start vs. not Head Start—and then extend this setup to the more general multi-valued treatment setting. Additional technical details are deferred to Appendix A.

[3] Family risk in HSIS is based on the sum of five variables: "(1) whether the household received food stamps or TANF in Fall 2002; (2) if neither parent was a high school graduate; (3) if neither parent is working; (4) if the mother was a teen mother; (5) and if the mother is a single mother" (Puma et al., 2010b).

4.1 Overview of Bayesian principal stratification

Following Neyman (1990) and Rubin (1974), we set up our problem using the potential outcomes notation. Thus, the causal effects of interest are defined regardless of the mode of inference. With this setup, we explore two common inferential approaches: moment-based and model-based. In the moment-based approach, the idea is to equate the causal quantities of interest with population moments, and then introduce identifying assumptions to create valid moment estimators. In this setting, a parameter is said to be point-identified if the moment equations and identifying assumptions yield a single estimate (see Zhang and Rubin, 2003, for relevant discussion). In the Bayesian model-based approach, by contrast, unobserved potential outcomes are treated as unknown parameters to be estimated given the model and the observed data. Importantly, identification issues are quite different from this perspective. In a Bayesian setting, proper prior distributions always yield proper posterior distributions. Thus, lack of identification results in regions of flatness of the posterior (Imbens and Rubin, 1997a), and identifying assumptions are not strictly necessary. Rather, introducing these assumptions sharpens the resulting inference.

Our primary approach in this paper is the parametric Bayesian paradigm, which has become widespread for principal stratification analysis (e.g., Hirano et al., 2000; Mattei et al., 2013). First, the Bayesian approach is attractive for causal inference with potential outcomes, which is essentially a missing data problem. Second, as Imbens and Rubin (1997b) discuss, parsimoniously parameterized models can often lead to better practical performance (in the sense of lower root-mean-squared-error) than corresponding moment-based approaches. Finally, we face a range of real-world complications in the HSIS example: missing data and study attrition; stratified randomization across many, small Head Start centers; and a mix of child- and center-level covariates. Addressing these issues is natural in a full Bayesian model but would be quite difficult with moment-based approaches.

At the same time, we still find it useful to articulate the assumptions necessary for a moment-based analysis. First, while hierarchical Bayesian modeling is a powerful inferential tool, it is often difficult to determine what "drives" such models in practice. Indeed, Cox and Donnelly (2011, p. 96) warn that "if an issue can be addressed nonparametrically then it will often be better to tackle it parametrically; however, if it cannot be resolved nonparametrically then it is usually dangerous to resolve it parametrically." We therefore believe it is useful to assess the level of danger we face. By thinking through the nonparametric approach, we show that the danger in our model is largely due to dependence on the Normality assumption.

4.2 Setup and ITT

We observe N children, N_1 of whom are randomized to receive the opportunity to enroll in Head Start, with treatment indicator $Z_i = 1$ for child i, and N_0 of whom are not, with $Z_i = 0$. We analyze the HSIS data as a stratified randomized evaluation, with child-level randomization conducted separately within each Head Start center.

In order to use the potential outcomes notation, we first make the standard Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1980), which states that the treatment assignment of one child does not affect the outcome of another child. Next, we define the relevant potential outcomes. First, let $D_i^{\text{obs}} \in \mathcal{D}$ denote the observed care setting for child i, where $\mathcal{D}$ is the set of possible care settings, and $D_i(z)$ is the care setting child i would have received if that child had been assigned to treatment condition z. Second, let $Y_i^{\text{obs}} \in \mathbb{R}$ denote the observed outcome of interest (e.g., PPVT), with corresponding potential outcomes, $Y_i(z)$. With this setup, $Y_i^{\text{obs}} = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ and $D_i^{\text{obs}} = Z_i D_i(1) + (1 - Z_i) D_i(0)$.

We now formalize the assumption that randomization is valid, which is sensible given that HSIS is a randomized experiment: (Imbens and Rubin, 2015):

Assumption R. *(Random assignment.)* Treatment assignment probabilities do not depend on the potential outcomes:

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1), D_i(0), D_i(1)) .$$

Finally, we define the Intent-to-Treat (ITT) estimand as

$$\text{ITT} = \frac{1}{N} \sum Y_i(1) - Y_i(0).$$

Under assumption R, we can estimate the ITT with the usual difference-in-means estimator. We note that our estimands of interest are defined for the finite sample of N children observed in HSIS, which is straightforward to estimate in the Bayesian paradigm (see Imbens and Rubin, 2015, for further discussion). However, since we also present moment-based results, we present all assumptions in terms of a super-population for convenience.

4.3 IV: $D_i^* \in \{\textbf{Head Start}, \textbf{Not Head Start}\}$

To introduce the overall approach, we briefly walk through the assumptions necessary to identify the Local Average Treatment Effect, following Angrist et al. (1996). Let D_i^* be a binary indicator for whether or not child i participated in Head Start in the first year. Define child i's compliance type, S_i^*, via the joint values $(D_i^*(0), D_i^*(1))$, as shown in Table 3a. For continuity with the next section, we refer to these compliance types by the more general term, principal strata, taking values $S_i^* \in \{\text{Always Head Start, Never Head Start, Complier, Defier}\}$. As usual, we define the LATE as the impact of randomization on the Compliers:

$$\text{LATE} = \frac{1}{N_{\text{c}}} \sum_{i:\ S_i^* = \text{c}} Y_i(1) - Y_i(0).$$

The two standard assumptions for IV are: (1) the "no defiers" assumption; and (2) the exclusion restrictions for Always Head Start and Never Head Start children.

Assumption IV-1. *(IV Monotonicity/No Defiers.)* There are no individuals with $\{D_i^*(0) = 1, D_i^*(1) = 0\}$.

The monotonicity assumption states that there are no children who would enroll in Head Start when denied access to the program but who would not enroll when explicitly offered a position. While such behavior is possible in other settings, it is unlikely in the context of HSIS, where enrolling in Head Start without an available position is already quite difficult.

Assumption IV-2. *(IV Exclusion Restrictions.)* For $S_i^* \in \{\text{Always Head Start}, \text{Never Head Start}\}$, $Y_i(0) = Y_i(1)$.

The exclusion restriction for Never Head Start children states that there is no effect of randomization on those children who would never enroll. While this is not always a plausible assumption (e.g., Jo and Stuart, 2009), in the context of Head Start, there is no reason to expect that turning down the offer of enrollment will have any effect on test scores. The exclusion restriction for Always Head Start children states that there is no effect of randomization on those children who would enroll in Head Start regardless of random assignment. As Gibbs et al. (2011) argue, this exclusion restriction might not hold in practice. In particular, roughly half of Always Head Start children enroll in Head Start centers other than the center of random assignment. If these alternative centers systematically differ from centers of random assignment, then the exclusion restriction might not hold for this group.

From a moment-based perspective, Assumptions IV-1 and IV-2 are necessary to identify the LATE, as we discuss in Section 6.2 (Angrist et al., 1996). From a Bayesian model-based perspective, these assumptions are not strictly necessary for inference. Therefore, it is possible to assess these assumptions by relaxing them in the model (e.g., Imbens and Rubin, 1997a; Hirano et al., 2000; Mattei et al., 2013). As these questions are not central to our main substantive point, however, we do not explore them further here.

4.4 Principal Stratification: $D_i \in \{\textbf{Head Start}, \textbf{Other Center}, \textbf{Home}\}$

The IV approach allows us to estimate the impact of Head Start among Compliers. However, we wish to estimate differential impacts for children within this group. Our inferential goal is to divide the overall LATE into one LATE for those who would otherwise receive care in another non-Head Start center and a second LATE for those who would otherwise receive care in a home-based setting. To do so, we disaggregate the binary indicator, D_i^*, to three levels: $D_i \in \{\text{Head Start}, \text{Other Center}, \text{Home}\}$. We also disaggregate the set of three standard compliance types, $\mathcal{S}^*$, into a more complete set of principal strata, $\mathcal{S}$. Table 3b shows the nine possible combinations of care types under both treatment and control. Column headings correspond to the type of care each child would experience if assigned to the control condition; row headings correspond to the type of care each would experience if assigned to the treatment condition.

As in the standard IV case, we make two key types of assumptions: monotonicity assumptions and exclusion restrictions. The standard monotonicity assumption from the IV setting becomes a statement about four strata rather than just one. We break this statement into two parts.

Assumption PS-1a. *(PS Monotonicity/No Defiers.)* There are no individuals with $\{D_i(0) = \text{HS}, D_i(1) = \text{Center}\}$ or $\{D_i(0) = \text{HS}, D_i(1) = \text{Home}\}$.

Assumption PS-1a states that there are no children who would take up Head Start under assignment to control but not under assignment to treatment. Therefore, strata A and C in Table 3b do not exist. This is a natural extension of Assumption IV-1 to multi-valued D. As with Defiers in the IV setup, these two types of Defiers are unlikely to exist in HSIS.

Assumption PS-1b. *(Irrelevant Alternatives.)* There are no individuals with $\{D_i(0) = \text{Center}, D_i(1) = \text{Home}\}$ or $\{D_i(0) = \text{Home}, D_i(1) = \text{Center}\}$.

Assumption PS-1b states that the Head Start offer does not change the care setting for families choosing between non-Head Start options. Therefore, strata B and D in Table 3b do not exist. Walters (2015) motivates this assumption with a revealed preference argument: since the availability of non-Head Start preschool is not affected by a Head Start offer, preferences among non-Head Start care options should not be affected either. While this is an unverifiable assumption, it is likely that, if such families do exist, they make up only a very small fraction of the overall population.

This yields five possible principal strata: Always Head Start (ahs), Always Center (ac), Always Home (ah), Center Complier (cc), and Home Complier (hc). As in the IV case, we can naturally make exclusion restrictions for principal strata unaffected by randomization. In particular we assume zero treatment effect for the Always Head Start, Always Center, and Always Home strata.

Assumption PS-2. *(PS Exclusion Restrictions.)* For $S_i \in \{\text{Always Head Start, Always Center, Always Home}\}$, $Y_i(0) = Y_i(1)$.

The exclusion restriction for Always Head Start children here is identical to the exclusion restriction for Always Head Start children in the IV case. The exclusion restriction for Never Head Start children in the IV case directly implies the exclusion restrictions for Always Center and Always Home children here.

The remaining strata are Center Compliers and Home Compliers. Our goal is to estimate the impacts of randomization for these groups, which are the effects of receiving Head Start versus receiving other center-based care and home-based care, respectively:

$$
\begin{aligned}
LATE_{\text{cc}} &= \frac{1}{N_{\text{cc}}} \sum_{i:\ S_i = \text{cc}} Y_i(1) - Y_i(0) \\
LATE_{\text{hc}} &= \frac{1}{N_{\text{hc}}} \sum_{i:\ S_i = \text{hc}} Y_i(1) - Y_i(0).
\end{aligned}
$$

As with the overall LATE, these are local effects since they are only defined for specific subgroups. In other words, we cannot interpret the difference between $LATE_{\text{cc}}$ and $LATE_{\text{hc}}$ as the causal effect of other center-based care versus home care—these two subgroups are not the same children. They differ across a range of unobserved and observed characteristics, such as child pre-test scores and family characteristics.

Finally, the overall LATE is a weighted average of these two estimands:

$$LATE = \frac{\pi_{\text{cc}}}{\pi_{\text{cc}} + \pi_{\text{hc}}} LATE_{\text{cc}} + \frac{\pi_{\text{hc}}}{\pi_{\text{cc}} + \pi_{\text{hc}}} LATE_{\text{hc}}$$

where π_s denotes the proportion of children in stratum s.

5 Describing Principal Strata

In most subgroup analyses, the groups themselves are known and fixed. For example, we can easily estimate the differential impact of Head Start for boys and girls: after collecting baseline data, each child's gender is known. While principal strata are well-defined subgroups, just like three and four year olds, we cannot directly observe subgroup membership for all children.

Fortunately, we can extend some results from the IV case to provide useful descriptions of the principal strata themselves. In particular, we can non-parametrically identify the overall distribution of principal strata as well as the distribution of covariates within each stratum. We could therefore use moment-based methods to estimate these distributions. However, as discussed in Section 4.1, we instead use a Bayesian model-based approach, which allows us to address important study complications. Unsurprisingly, we find that these principal strata indeed differ across observed characteristics and that this variation is consistent with intuition and results in the early childhood literature.

The Appendix gives further details for the results we present below along with proofs of all the lemmas.

5.1 Overall Distribution of Principal Strata

Extending the standard results from the IV case (Angrist et al., 1996), we can estimate the overall size of each principal stratum.

Lemma 1 (Distribution of Principal Strata)**.** *Under Assumptions R, PS-1a, and PS-1b, the distribution of principal strata,* $\pi_s \equiv \mathbb{P}\{S_i = s\}$*, is non-parametrically identified for all* s*.*

For intuition on Lemma 1, it is useful to see the analogue in the IV setting: we first estimate the proportion of Always Head Start children in the control group and Never Head Start children in the treatment group, and then subtract to estimate the proportion of Compliers. Table 4 shows point estimates for the distribution of principal strata in the sample. Roughly one-third of all children are non-compliers of various types; each non-complier stratum is around 10 percent of the overall sample. The remaining two-thirds are split between the two Complier groups; Home Compliers total around 70 percent of all Compliers.

5.2 Using Covariates to Predict Stratum Membership

Since HSIS is a randomized experiment, we can examine the distribution of principal strata for specific subgroups, such as for all boys in the sample. Following Hill et al. (2002), we define the *principal score* as $\pi_{s|\mathbf{x}} \equiv \mathbb{P}(S_i = s \mid \mathbf{X}_i = \mathbf{x})$, the probability that a child belongs to principal stratum s given that child's observed covariates (see also Abadie, 2003; Jo and Stuart, 2009). Note that this is a simple generalization of modeling the "first stage" in the standard IV setting as a function of the covariates (e.g., Angrist, 2004).[4]

For HSIS, we estimate the principal score using multinomial logistic regression and a simple data augmentation procedure.[5] Figure 1 shows the resulting logistic regression coefficients for select covariates that are predictive of being a Center-based vs. Home-based Complier. We discuss these results below.

5.3 Distribution of Covariates by Principal Stratum

We can also estimate the distribution of covariates for each principal stratum.

Lemma 2 (Distribution of Covariates by Principal Stratum)**.** *Under Assumptions R, PS-1a, and PS-1b,* $\mathbb{P}\{\mathbf{X}_i = \mathbf{x} \mid S_i = s\}$ *is non-parametrically identified for all* s*.*

This lemma is a simple extension of the comparable IV result in Abadie (2003), and allows us to make concrete observations about otherwise unobservable groups (see also Angrist and Pischke, 2008; Frumento et al., 2012). Table 6 shows the means for select covariates for each stratum; Figure 2 separately shows the means for pre-test score by principal stratum. There are key differences in observable characteristics across the latent groups. Columns 1–3 on Table 6 show variation in pre-treatment covariates across the different types of non-compliers. Overall, these results suggest that children who always enroll in a non-Head Start center-based setting outperform their counterparts who would always be in Head Start or in a home-based setting. For example, as shown in Figure 2, Always Center-based children strongly outperform Always Head Start and Always Home-based children on the PPVT pre-test. Other covariates also sensibly predict differences among the non-complier types. For example, Always Center children are much more likely to live in a state that has state-funded preschool than Always Home children. In general, this ordering is consistent with the selection results from Deming (2009), who finds that families of children in non-Head Start preschools have higher income and maternal education than families of children in Head Start or in no preschool.

[4] Unlike the usual first stage model, $\mathbb{P}\{D^{*,\text{obs}} \mid X_i = \mathbf{x}\}$, the principal score is vector-valued, since S_i is discrete rather than binary.

[5] This approach improves on simpler versions of this model fit by Walters (2015) and Zhai et al. (2014). Walters (2015) effectively estimates the share of Center-based Compliers and Home-based Compliers for each Head Start center, doing so via two separate logistic regressions, rather than via multinomial logistic regression. Zhai et al. (2014) estimate a multinomial logistic regression using covariates to predict $D(0)$ rather than stratum membership, therefore conflating Always Center-based children and Center Compliers under control and conflating Always Home-based children and Home Compliers under control.

We can compare our two complier groups by examining columns 4 and 5 of Table 6, which are the complement to the logistic regression coefficients in Figure 1. Consistent with research that has found that parents typically prefer center-based care for four-year olds (e.g., Huston et al., 2002; Rose and Elicker, 2010), roughly 60 percent of Home Compliers are three years old, compared to only 45 percent of Center Compliers. We also find that Home Compliers enter the study with lower pre-academic skills. Home Compliers exhibit lower PPVT performance at the beginning of the study and are more likely to be in the bottom third of PPVT performance compared to Center Compliers. Home Compliers are additionally more likely to have a mother with less than a high school education. As above, Center Compliers are more likely to live in states that, during the time of the HSIS, provided state-funded pre-kindergarten. Note that we do not find meaningful differences between these two groups based on race or ethnicity or based on Dual Language Learner status.

Overall, these differences in covariate means by principal stratum underscore that children in different principal strata do, indeed, differ in terms of their baseline characteristics. Therefore, while estimates of causal effects *within* each principal stratum are valid, comparisons *between* principal strata are descriptive rather than causal, in the same way that comparing treatment effects for males and females is descriptive rather than causal. In other words, differential impacts across strata could also be due to differences in observed or unobserved characteristics other than care type. See Gallop et al. (2009) for a discussion of using principal stratification for mediation analysis, which generally requires much stronger assumptions than those presented here.

6 Overview of Identification and Estimation

This section provides an overview of the identification and estimation strategies used in this paper. Interested readers can find greater detail in the Appendix, which gives an in-depth discussion of possible identification approaches, our hierarchical Bayesian estimation procedure, robustness to different parametric assumptions, and other technical information. Conversely, readers can skip to Section 7 for a discussion of the results.

6.1 Identification

The identification strategy rests on the idea that we can identify the outcome distributions for each principal stratum. This builds on earlier work in the IV case from Imbens and Rubin (1997b) and Abadie (2003). We provide a brief sketch of the idea here. The Appendix provides additional discussion of identification in principal stratification models (see also Zhang et al., 2009).

To illustrate the identification approach, first consider a standard subgroup analysis, for example, estimating the impact of Head Start for the subgroup of boys. Formally, we can achieve this in two distinct steps. The first step is to identify the distribution of outcomes for boys in the treatment group, which we denote $g_{\text{boys}\,1}(y)$, and the corresponding distribution of outcomes for boys in the control group, which we denote $g_{\text{boys}\,0}(y)$. Since HSIS is a randomized experiment, and since we directly observe which children are boys, we can non-parametrically identify both $g_{\text{boys}\,0}(y)$ and $g_{\text{boys}\,1}(y)$ from the corresponding sample (e.g., via kernel density estimation). We can then obtain the average impact of Head Start on boys by comparing the means of the two distributions. While not necessarily practical, this is nonetheless a valid procedure for identifying an average treatment effect for a subgroup.

6.1.1 Instrumental Variables

While we directly observe gender, we do not directly observe compliance type for all children. We therefore must adopt a different approach for estimating the outcome distributions by compliance type. For illustration, we again begin with the standard IV set up for non-compliance, where we compare Head Start versus not Head Start:

- **Always Head Start and Never Head Start.** Under monotonicity, we know that any children in the control group who enroll in Head Start must be Always Head Start children. As a result, we can directly estimate the outcome distribution for the Always Head Start subgroup under control, $g_{\text{ahs}\,0}(y)$. Since we assume that there is no treatment effect for this group (i.e., that the exclusion restriction holds for Always Head Start children), then $g_{\text{ahs}\,1}(y) = g_{\text{ahs}\,0}(y) = g_{\text{ahs}}(y)$. We can repeat this approach for Never Head Start children in the treatment group, which yields $g_{\text{nhs}\,1}(y) = g_{\text{nhs}\,0}(y) = g_{\text{nhs}}(y)$.

- **Compliers.** We must take a different approach for the Compliers. First, we cannot directly observe which children are Compliers. Second, since we are interested in the LATE, we can no longer assume that Compliers have the same outcome distribution under treatment and control. The key insight is to focus on the relationship between the observed treatment and the unobserved compliance type; Table 5a shows these relationships for the IV case. For example, children in the control group who do not enroll in Head Start are either Compliers or Never Head Start children. In other words, the observed outcome distribution for these children is a mixture of $g_{\text{nhs}}(y)$ and $g_{\text{co}\,0}(y)$. Formally:

$$f_{00}(y) = \frac{\pi_{\text{nhs}}}{\pi_{\text{nhs}} + \pi_{\text{co}}} g_{\text{nhs}}(y) + \frac{\pi_{\text{co}}}{\pi_{\text{nhs}} + \pi_{\text{co}}} g_{\text{co}\,0}(y), \tag{1}$$

  where $f_{zd}(y)$ is the observed outcome distribution for children with treatment assignment $Z_i = z$ and treatment received $D_i^* = d$. For example, $f_{00}(y)$ is the observed outcome distribution for children assigned to the control condition who do not experience Head Start. Since we can directly observe $f_{00}(y)$, π_{nhs}, π_{co}, and $g_{\text{nhs}}(y)$, we can re-arrange terms to identify $g_{\text{co}\,0}(y)$, the outcome distribution for Complier children in the control group. We can repeat this with the mixture of Always Head Start and Compliers under treatment to obtain $g_{\text{co}\,1}(y)$. Therefore, we can non-parametrically identify both $g_{\text{co}\,0}(y)$ and $g_{\text{co}\,1}(y)$, even though we cannot observe these distributions directly. See Imbens and Rubin (1997b) for additional discussion.

Once we have all the outcome distributions, we can immediately obtain the average outcomes by principal stratum, μ_{sz}, and finally obtain $LATE = \mu_{\text{co}\,1} - \mu_{\text{co}\,0}$. Also see Kling et al. (2007) for an example in which the Complier means are substantively meaningful in their own right. More generally, Abadie (2003) shows that we can use this approach to identify a broad range of features by compliance type, including covariate distributions.

6.1.2 Principal Stratification

We now extend the argument from the IV case to identify the outcome distributions for our principal strata of interest. We again have observed mixtures, as shown in Table 5b.

- **Always Head Start, Always Center-based, and Always Home-based.** Just as with the Always Head Start and Never Head Start groups, we directly observe the outcome distributions for the Always Head Start, Always Center-based, and Always Home-based strata. For example, we directly observe the Always Home-based children under treatment and can therefore non-parametrically identify $g_{\text{ah}\,1}(y)$. Since we assume that there is no impact of randomization on this group, $g_{\text{ah}\,1}(y) = g_{\text{ah}\,0}(y) = g_{\text{ah}}(y)$. We repeat this for the Always Head Start and Always Center-based strata, yielding non-parametric identification for $g_{\text{ahs}}(y)$, $g_{\text{ac}}(y)$, and $g_{\text{ah}}(y)$.

- **Center-based Compliers (control) and Home-based Compliers (control).** As in the IV case, we cannot directly observe the outcome distributions for Center-based Compliers and Home-based Compliers and must instead identify these distributions indirectly. We begin with the outcome distribution for Home-based Compliers under control, $g_{\text{hc}\,0}(y)$. Analogous to Equation (1), the outcome distribution for control group children in home-based care is a mixture of $g_{\text{ah}}(y)$ and $g_{\text{hc}\,0}(y)$:

$$f_{0\,\text{Home}}(y) = \frac{\pi_{\text{ah}}}{\pi_{\text{ah}} + \pi_{\text{hc}}}\, g_{\text{ah}}(y) + \frac{\pi_{\text{hc}}}{\pi_{\text{ah}} + \pi_{\text{hc}}}\, g_{\text{hc}\,0}(y). \tag{2}$$

  where we have previously identified $g_{\text{ah}}(y)$. Similarly, we re-arrange terms to non-parametrically identify $g_{\text{hc}\,0}(y)$ and repeat this procedure for the Center-based Compliers under control, $g_{\text{cc}\,0}(y)$.

- **Center-based Compliers (treated) and Home-based Compliers (treated).** Identifying the corresponding Complier distributions under treatment requires additional steps. As in the IV case, we can reduce the problem to estimating a mixture of two types:

$$f^{*}_{1\,\text{HS}}(y) = \frac{\pi_{\text{cc}}}{\pi_{\text{cc}} + \pi_{\text{hc}}}\, g_{\text{cc}\,1}(y) + \frac{\pi_{\text{hc}}}{\pi_{\text{cc}} + \pi_{\text{hc}}}\, g_{\text{hc}\,1}(y). \tag{3}$$

  where $f^{*}_{1\,\text{HS}}(y)$ is the observed outcome distribution after "backing out" the Always Head Start outcome distribution. Unlike the IV case, however, neither mixture component is known, which leads to a two-component finite mixture. Without additional assumptions, the component densities, $g_{\mathrm{cc}\,1}(y)$ and $g_{\mathrm{hc}\,1}(y)$, are not identified.

Therefore, the key inferential challenge, at least implicitly, is estimating the parameters of a two-component finite mixture. Once we obtain the relevant component means, $\mu_{\mathrm{cc}\,1}$ and $\mu_{\mathrm{hc}\,1}$, we can then estimate $LATE_{\mathrm{cc}} = \mu_{\mathrm{cc}\,1} - \mu_{\mathrm{cc}\,0}$ and $LATE_{\mathrm{hc}} = \mu_{\mathrm{hc}\,1} - \mu_{\mathrm{hc}\,0}$.

There are many possible approaches to disentangle the finite mixture model. Since we adopt a Bayesian parametric framework here, it is natural to assume that the component densities, $g_{\mathrm{cc}\,1}(y)$ and $g_{\mathrm{hc}\,1}(y)$, follow a parametric distribution, namely Normality. In a classic result, Pearson (1894) showed that the component parameters are all identified under this assumption. Similar results hold for a broad class of parametric models (Frühwirth-Schnatter, 2006) and for distributions with shape restrictions, such as symmetry (Bordes et al., 2006; Hunter et al., 2007). Note that, as we discuss in the next section, our model imposes the Normality assumption on the outcome residuals (i.e., conditional on covariates) rather than on the marginal outcome distributions.

Finally, it is useful to briefly review some alternative strategies that leverage auxiliary covariates to disentangle the finite mixture model (Joffe et al., 2007). First, researchers cam assume that that, conditional on covariates, stratum membership is independent of potential outcomes, an assumption known as principal ignorability. This can be a sensible assumption in some settings (e.g., Hill et al., 2002; Schochet and Burghardt, 2007; Scott-Clayton and Minaya, 2014), but seems somewhat implausible here, as we do not observe critical variables like parental preference for care type prior to randomization. Second, researchers can restrict the relationship between a special covariate and the outcome; for example, assuming that the treatment effect does not vary across site (Raudenbush et al., 2012). While many such restrictions are possible (e.g., Jo, 2002; Ding et al., 2011; Mealli and Pacini, 2013), there is no clear candidate for such a special covariate in HSIS, nor is it plausible to assume that the treatment effect is constant across Head Start centers. Finally, see Hall and Zhou (2003) and Mealli and Pacini (2013) for assumptions when there are multiple, independent outcomes.

6.2 Estimation

We now turn to model-based estimation. In practice, we could estimate the full parametric model from either a likelihood or Bayesian perspective. Indeed, some prominent applications of model-based principal stratification utilize a direct likelihood approach (e.g., Zhang et al., 2009; Frumento et al., 2012). This approach is quite flexible and allows for straightforward comparisons between different models. It is especially attractive when specifying prior distributions is not desirable. An important feature of the Head Start data, however, is the multilevel structure of children nested within Head Start centers. Incorporating this structure is immediate with a Bayesian approach but can prove quite complex in a likelihood setting. In addition, accounting for uncertainty in the parameter estimates is natural with a Bayesian approach but can be more involved with a direct likelihood approach (see, for example, Frumento et al., 2016). While we use a Bayesian estimation approach, we would expect quite similar results using either method.

6.2.1 Sketch of Data Augmentation

To develop intuition, we first give a high-level sketch of a data augmentation procedure for estimating the parameters of interest. To focus on the core estimation problem, we initially ignore important complications, returning to them below. The key idea is to alternate between (1) estimate the vector of model parameters, θ, given stratum membership, S, and (2) imputing each child's principal stratum membership, S, given θ. Beginning with an initial guess of principal stratum membership for each child:

- **Step 1: Given stratum membership, estimate model parameters.** We estimate model parameters via two sub-models.
  - *Step 1A: Outcome sub-model,* $g_{sz|\mathbf{x}}(y)$. First, we estimate the regression of Y^{obs} on $\mathbf{X}$ and Z within each principal stratum, S. The critical assumption is that the residuals follow a Normal distribution.
  - *Step 1B: Principal score sub-model,* $\pi_{s|\mathbf{x}}$. Second, we estimate a multinomial logistic regression predicting S given $\mathbf{X}$.
- **Step 2: Given model parameters, predict stratum membership.** Given the outcome sub-model, $g_{sz|\mathbf{x}}(y)$, and principal score sub-model, $\pi_{s|\mathbf{x}}$, we can estimate the probability of stratum membership via Bayes' Rule. For example, if we observe a child in the control group who is in home-based care, the child's probability of being a Home Complier is:

  $$\mathbb{P}\{S_i = \text{hc} \mid \text{data}, \theta\} = \frac{\pi_{\text{hc}|\mathbf{x}} \cdot g_{\text{hc}\,0|\mathbf{x}}(y)}{\pi_{\text{hc}|\mathbf{x}} \cdot g_{\text{hc}\,0|\mathbf{x}}(y) + \pi_{\text{ah}|\mathbf{x}} \cdot g_{\text{ah}|\mathbf{x}}(y)}.$$

  We then flip a weighted coin to predict S_i for that child. By contrast, if we observe a child in the treatment group who is in home-based care, the child must be in the Always Home-based stratum. So $\mathbb{P}\{S_i = \text{ah} \mid \text{data}, \theta\} = 1$.

6.2.2 Model details

The actual model is considerably more complex. We highlight key issues here and defer additional technical details to the appendix. First, the outcome models by principal stratum are:

$$\begin{aligned}
y_i^{\text{obs}} \mid (S_i = \text{ahs}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{ahs}} + \beta_{\text{ahs}}\mathbf{x}_i + \psi_{j[i]}, \sigma^2_{\text{ahs}}\right)\\
y_i^{\text{obs}} \mid (S_i = \text{ac}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{ac}} + \beta_{\text{ac}}\mathbf{x}_i + \psi_{j[i]}, \sigma^2_{\text{ac}}\right)\\
y_i^{\text{obs}} \mid (S_i = \text{ah}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{ah}} + \beta_{\text{ah}}\mathbf{x}_i + \psi_{j[i]}, \sigma^2_{\text{ah}}\right)\\
y_i^{\text{obs}} \mid (S_i = \text{cc}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{cc}} + \beta_{\text{cc}}\mathbf{x}_i + \psi_{j[i]} + \tau_{\text{cc}}z_i + \omega_{j[i],\text{cc}}z_i, \sigma^2_{\text{cc},z}\right)\\
y_i^{\text{obs}} \mid (S_i = \text{hc}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{hc}} + \beta_{\text{hc}}\mathbf{x}_i + \psi_{j[i]} + \tau_{\text{hc}}z_i + \omega_{j[i],\text{hc}}z_i, \sigma^2_{\text{hc},z}\right),
\end{aligned}$$

where $j[i]$ denotes the site j corresponding to child i. Within each stratum, this is essentially a varying intercept/varying slope model. To improve the stability of the model, the variance terms for the two complier groups under treatment are constrained to be equal, $\sigma^2_{\text{cc}\,1} = \sigma^2_{\text{hc}\,1}$.[6] Given small sample sizes within each site, the random effects for site, $\{\psi_j\}$, are constrained to be equal across principal strata, although the treatment effects are allowed to differ. The site-level estimates follow a multivariate Normal distribution:

$$\begin{pmatrix} \psi_j \\ \omega_{j,\text{cc}} \\ \omega_{j,\text{hc}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \gamma^{ctr}\mathbf{w}_j \\ 0 \\ 0 \end{pmatrix}, \Sigma_y \right)$$

where $\mathbf{w}_j$ is a vector of site-level covariates and Σ_y is an unconstrained covariance matrix. We include the proportion assigned to treatment, $\bar{z}_j$, as a site-level predictor in order to account for differing proportions randomized to treatment by site (see Bafumi and Gelman, 2006; Raudenbush, 2015).

We also introduce a multilevel structure in the multinomial logistic regression model:

$$\begin{aligned} \mathbb{P}(S_i = s \mid \theta, \mathbf{x}_i) &= \frac{\exp(\gamma_{s,j[i]} + \delta_s' \mathbf{x}_i)}{\sum_{s=1}^{K} \exp(\gamma_{s,j[i]} + \delta_s' \mathbf{x}_i)} \\ \gamma_{s,j} &\sim \mathcal{N}(\mu_{\gamma,s} + \delta_s^{ctr}\mathbf{w}_j, \eta^2_{\gamma,s}) \end{aligned}$$

where the site-level random effects are independent across strata. See the appendix for additional details.

Three additional points are worth noting. First, as discussed in Section 3.2, there is considerable missingness in HSIS, especially in the outcomes. We address this by assuming that outcomes are Missing at Random (MAR) (Rubin, 1976),

$$\mathbb{P}\{M_i \mid Y_i, \mathbf{X}_i, Z_i, D_i^{obs}\} = \mathbb{P}\{M_i \mid \mathbf{X}_i, Z_i, D_i^{obs}\},$$

where M_i is an indicator for missing outcome. In other words, given covariates, treatment assignment, and observed child care setting, missing outcomes are just as likely to be low test scores as high test scores. While we address alternative assumptions in the appendix, MAR is at least plausible for HSIS, since the data collection procedures depended heavily on the child's actual care setting. Although implicit, this is also the assumption behind the nonresponse adjustment in the official HSIS report (Puma et al., 2010a).

Second, as we discuss in Section 7.3, the treatment effect varies across observed covariates. Given the complexity of the base model, however, we report these treatment-by-covariate interactions one at a time. Since including multiple treatment-by-covariate interactions unsurprisingly yields poor

[6] Relaxing the constraint that $\sigma^2_{\text{cc}\,1} = \sigma^2_{\text{hc}\,1}$ gives comparable results but leads to worse model fit, since identification for these variance terms is rather weak. Alternatively, Imbens and Rubin (1997a) suggest modeling the variance based on treatment received rather than treatment assigned, which would lead to $\sigma^2_{\text{cc}\,1} = \sigma^2_{\text{hc}\,1} = \sigma^2_{\text{ahs}}$ in this context. While this is a stronger assumption than the equal variance case above, invoking this assumption reduces the number of unknown parameters in the mixture model by one. See also Griffin et al. (2008).

model convergence, the main results are from a model that excludes such interactions. Finally, we use standard reference priors throughout. See the appendix for additional details.

6.2.3 Computational details

While this data augmentation procedure helps to build intuition for estimation, convergence of the algorithm can be slow in practice. Instead, we estimate this model via Stan, a Bayesian programming language that implements a variant of Hamiltonian Monte Carlo (HMC; Stan Development Team, 2014; Hoffman and Gelman, 2014). Unlike, say, a classic Gibbs sampler, HMC-based samplers explore the space of the (log) posterior far more efficiently than more standard Markov chain Monte Carlo approaches, dramatically increasing the effective sample size of the same number of draws (Hoffman and Gelman, 2014). One drawback of the HMC approach is that the log-posterior must have globally smooth gradients. As a result, Stan/HMC cannot incorporate discrete latent parameters, such as indicators for principal stratum membership that would be standard in a data augmentation scheme. Stan sidesteps this issue by maximizing the observed data log-posterior rather than the complete data log-posterior. While it is possible to couple a data augmentation Gibbs step with a bespoke HMC sampler, doing so would lose many of Stan's key advantages, including optimized C++ code and a powerful, flexible programming language. In the end, it is unlikely that this project would have been feasible without the development of Stan.

Each model was run with five separate chains with 500 "warm up" draws and 500 posterior draws. We assess model convergence in the usual way via traceplots, via Gelman-Rubin $\widehat{R}$ statistics at or near 1, and via measures of the effective sample size from each chain. All models reported here showed excellent convergence for parameters of interest. As with all hierarchical models, some hyperparameters were poorly estimated; we do not report those.

7 Results

We now summarize results for the Intent-to-Treat, Instrumental Variable, and Principal Stratification models, beginning with impacts in the Head Start year. We then briefly explore impacts after the first year as well as additional impact heterogeneity, including distributional treatment effects. Finally, we report sensitivity and robustness checks.

7.1 Impacts in the Head Start Year

The first row of Table 7 shows the ITT estimate, the impact of opportunity to enroll in Head Start, on PPVT in effect size units (i.e., effects scaled by the SD of the control group). Consistent with the original Head Start results (Puma et al., 2010a), we find that the overall impact of randomization to treatment is +0.14 in the Head Start year (posterior median). There is strong evidence that this impact is greater than zero.

In general, the results we present here give much stronger statistical evidence that the impacts are positive than the evidence presented in Puma et al. (2010a). Multiple factors contribute to these differences. First, unlike Puma et al. (2010a), we pool the three- and four-year-old cohorts, which roughly doubles the sample size. Second, unlike Puma et al. (2010a), we control for Head Start center of random assignment in the outcome model, which improves precision. Finally, we do not use the HSIS weights, which were created to generalize the experimental results to a particular population of Head Start children. As we estimate impacts for the finite sample of children in HSIS, the corresponding standard errors are smaller. See Bloom and Weiland (2014) for additional discussion.

The second row of Table 7 shows the corresponding LATE estimate from the IV model. Among Compliers, the impact of enrolling in Head Start on PPVT is +0.18. This estimate is nearly identical to that of Bloom and Weiland (2014), who conduct a similar analysis. As with the ITT, there is strong evidence that this impact is positive.[7] This effect is comparable to the average effects of early childhood education programs reported in a recent meta-analysis (effect size of +0.21; Duncan and Magnuson, 2013) and represents approximately one-quarter of the Black-White test score gap at the end of kindergarten (Fryer and Levitt, 2004).

The last three rows of Table 7 show the principal stratification results from the full model. For Home Compliers, we find a treatment effect of +0.23 on PPVT, with strong evidence that these impacts are greater than zero. This is much larger than the ITT effect. For Center Compliers, however, we find an effect of zero. Because we jointly estimate $LATE_{\text{hc}}$ and $LATE_{\text{cc}}$, we can calculate that $\mathbb{P}\{LATE_{\text{hc}} > LATE_{\text{cc}}\} = 0.99$. As we discussed above, this is a descriptive comparison—like claiming that the treatment effect is larger for boys than girls—but it nonetheless shows that impacts for these two latent groups are meaningfully different.

A useful check is to compare the implied LATE and ITT estimates from the principal stratification model with the corresponding estimates from the IV and ITT models, respectively. In particular, the implied LATE is +0.16, which is quite close to the IV model estimate of +0.18; the implied ITT is +0.11, again close to the ITT model estimate of +0.14. This similarity is reassuring given the additional flexibility and complexity of the principal stratification model.

Another useful check is to compare our results to those of Kline and Walters (2015), who use a structural model to estimate a range of different treatment effects for the HSIS data, including $LATE_{\text{hc}}$ and $LATE_{\text{cc}}$. Identification in the Kline and Walters (2015) paper comes from two main sources: (1) assuming that the choice of a child's care setting follows a multinomial Probit discrete choice model (i.e., that the latent choice utilities follow a multivariate Normal distribution), and (2) assuming that there is no interaction between covariates and Z. First, our multinomial logistic regression model is analogous to their multinomial Probit model, although our modeling choice is not critical for identification. Second, our assumption of Normality on the residuals broadly takes the place of their assumption of no interaction between covariates and Z: both place restrictions on the heterogeneity of the outcome distributions. Thus, while our approaches are quite different in formulation (see Mealli and Pacini, 2008, for a comparison of selection models and principal

[7] Note that this estimate differs from the usual Wald estimator for IV, $\frac{ITT}{\pi_c} = \frac{0.14}{0.7} = 0.20$. This is primarily due to the multi-site randomization and differences in compliance rates across Head Start centers. See Raudenbush et al. (2012) and Reardon and Raudenbush (2013) for further discussion of this issue.

stratification), the underlying assumptions are similar in spirit. It is therefore reassuring that Kline and Walters (2015) also find the same overall pattern of effects, with positive and significant impacts for Home Compliers and negligible impacts for Center Compliers. While their point estimate for $LATE_{\text{hc}}$ is somewhat larger than ours (0.35 vs. 0.23), it appears as though this discrepancy is largely due to a different choice of outcome; Kline and Walters (2015) estimate impacts on an index of outcomes while we focus on PPVT alone.[8]

7.2 Impacts after the Head Start Year

A key feature of the HSIS design is that children in the three-year-old cohort control group were given access to the Head Start program in the second year of the study. In practice, nearly half of the control group took up the opportunity to enroll, with another 34 percent enrolling in other, non-Head Start center care during that year. Enrollment was similarly high for treatment group children: 64 percent enrolled in Head Start, with another 24 percent enrolling in other center care.[9] Therefore, by the second year of HSIS, the randomization only increased the probability of enrolling in Head Start by 16 percentage points and only increased the probability of enrolling in any center-based care setting by 6 percentage points.

There are several possible approaches to address this complication. First, we could expand the number of principal strata to allow for two years of enrollment in Head Start. However, this is impractical given the complexity of just modeling care setting in the first year. Another possibility is to re-define care setting to be $\mathcal{D} \in \{\text{Ever in Head Start}, \text{Home-based care}, \text{Center-based care}\}$. See, for example, Kline and Walters (2015). Consistent with the official report (Puma et al., 2010a), we focus on the setting in which the child was cared for in the first year of the intervention, even for outcomes collected in subsequent years. We believe that this is a sensible definition, as the randomization encourages participation in Head Start in the first year only. Nonetheless, simply pooling cohorts after the Head Start year does not yield easily interpretable results.

Following Puma et al. (2010a), we therefore analyze the results separately by cohort to assess impacts after the Head Start year. Unfortunately, further dividing Center and Home Compliers into separate three- and four-year-old subgroups makes estimation more challenging. Sample sizes are relatively small. In addition, outcome missingness increases substantially over the course of the study, with roughly a quarter of all outcomes missing by the third year. Therefore, the cohort and subgroup results presented below should all be considered exploratory.

With this caveat in mind, Figure 3 shows the treatment effect on PPVT by cohort by assessment year for all Compliers, for Center Compliers, and Home Compliers.[10] Consistent with the official

[8] We can assess the influence of the outcome choice with a simple back-of-the-envelope calculation. Their estimate of the overall $LATE$, which is non-parametrically identified, is roughly 40 percent larger than ours (0.25 vs. 0.18); their estimate of $LATE_{\text{hc}}$ is roughly 50 percent larger than ours (0.35 vs. 0.23).

[9] For the three year old cohort, 22 percent of control group children and 14 percent of treatment group children do not have an observed care setting in the second year of the study. Reported percentages are among children with observed care type.

[10] These are the normative grades for a given cohort. Children who began the study as three-year-olds were able to gain access to Head Start in year 2 and then enrolled in kindergarten in year 3. The four-year-olds transitioned

HSIS results, we find a decline in the treatment effect as children age. Nonetheless, unlike in the official HSIS results, we find impacts that are positive and meaningfully different from zero by the time children are in 1st grade, with $LATE$ estimates of 0.09 and 0.14 for the three- and four-year-old cohorts, respectively. The effects for Home Compliers follow the same decline as for the overall Compliers, albeit with slightly larger point estimates and with less precision. By contrast, the impacts among Center Compliers are best described as noise around zero, though this null result could be due to the limited sample size. Note that the pooled main effects in Table 7 are the (weighted) average of the impacts on three-year-olds at age 3 and four-year-olds at age 4.

While we regard these results as exploratory, they nonetheless suggest that the impact of Head Start might indeed persist into early elementary school, even if the magnitudes are modest. In particular, Gibbs et al. (2011) argue that a key puzzle of the HSIS results is not that they decrease over time, but that they attenuate to zero as soon as children leave the program, much more rapidly than estimates based on quasi-experimental methods (e.g., Currie and Thomas, 1993; Deming, 2009). The results in Figure 3 show that the decline in treatment effects may not be nearly as rapid as in the reported topline results.

7.3 Subgroup and Quantile Treatment Effects

Several recent papers have explored variation in Head Start's impact across observed subgroups and across quantiles of the outcome distribution. Since Center and Home Compliers differ across a range of observed and unobserved characteristics, an important question is therefore the extent to which these differences explain the different impacts for the two Complier groups. Again, these estimates should be considered exploratory.

First, we turn to variation across subgroups defined by pre-treatment characteristics. Following Bloom and Weiland (2014) and Bitler et al. (2014), we focus on variation by (1) whether a child is in the bottom third of pre-test score by cohort; and (2) whether a child is a Dual-Language Learner (DLL). Table 8 shows the corresponding principal stratification estimates during the Head Start year. First, across all four subgroups, we observe the same pattern of positive, significant effects for Home Compliers and negligible effects for Center Compliers. While the smaller sample sizes limit statistical power, this consistency nonetheless bolsters the overall findings. Second, as in Bloom and Weiland (2014), we find larger Home Complier effects for children in the bottom third by pre-test score and also for DLL students. The effect for DLL students is especially striking, with an effect size of around +0.35 SD in the Head Start year, more than double the point estimate for non-DLL students. This suggests that, at least in terms of vocabulary development, there is substantial impact of Head Start relative to a home-based setting in which English is likely not spoken. See Bloom and Weiland (2014) for additional discussion.

Another likely source of impact variation is heterogeneity across the outcome distribution (Bitler et al., 2003). In a recent paper, Bitler et al. (2014) estimate distributional effects for Head Start

to kindergarten and then first grade in the second and third years of the study. Therefore, by year 3, all children, if following a standard educational trajectory, were in elementary school.

via quantile treatment effects, $G^{-1}_{\text{co}\,1}(q) - G^{-1}_{\text{co}\,0}(q)$, the difference between the qth quantiles of the outcome distributions for Compliers under treatment and control, respectively. The authors find that the impacts of Head Start on PPVT and other measures are largest at the bottom of the outcome distribution, both overall and among Compliers. As we discuss in Appendix B, we can leverage our framework both to replicate and to extend their results. Figure 4 shows the quantile treatment effect estimates for all Compliers, Center Compliers, and Home Compliers during the Head Start year. As expected, our estimates for all Compliers are very close to those of Bitler et al. (2014), showing large, positive effects at the bottom of the distribution of between +0.4 and +0.6 SD. The effects for Home Compliers are also positive and significant throughout, with larger effects at the bottom of the distribution. By contrast, the quantile treatment effects for Center Compliers are essentially zero across the entire distribution.

7.4 Sensitivity Checks

We conducted robustness checks for our main results of impacts in the Head Start year, which we briefly discuss here. First, we assess sensitivity to our handling of missing data and re-fit the principal stratification model using only observed outcomes, approximately 80 percent of the overall sample. Table 9 shows the resulting complete case estimates, which are essentially unchanged from the full version. Second, as we discuss in Section 6, the Normality assumption plays a critical role in both identification and estimation. Table 9 shows the same model using a heavy-tailed Student t_7 distribution rather than a Gaussian. Again, the results are consistent.

Finally, following Rubin et al. (1984) and Gelman et al. (2013), we use posterior predictive checks to assess the fit of our full model to the observed data. Formally, let y be the observed data and θ be the parameter vector. Define y^{rep} as the replicated data that could have been observed if the study were replicated with the same model and the same value of θ that produced y. We can estimate the distribution of y^{rep} via the posterior predictive distribution,

$$p(y^{\text{rep}} \mid y) = \int p(y^{\text{rep}} \mid \theta) p(\theta \mid y) d\theta.$$

The intuition is to assess whether the replicated data produced from the model are similar to the observed data. In Appendix B.7, we assess this similarity in two ways. First, we visually inspect the observed and replicated data sets (see Appendix Figure A2). Second, we compute posterior predictive p-values following a similar approach in Barnard et al. (2003) and Mattei et al. (2013) (see Appendix Table A1). Neither approach yields evidence that the model is a poor fit to the data.

8 Discussion

Our primary contribution is to develop a framework for estimating impact variation by alternative care setting and to apply this framework to the Head Start Impact Study. In particular, we find positive and meaningful impacts on key outcomes among Home-based Compliers, those children who would enroll in Head Start under treatment and who would otherwise be in home-based care. By contrast, we find no meaningful effects among Center-based Compliers, those children who would otherwise receive non-Head Start center care.

In doing so, we present a much more nuanced view of Head Start's impact than the topline experimental results indicate. We also refute sweeping generalizations made about Head Start, such as "Head Start does not improve the school readiness of children from low-income families" (Whitehurst, 2013a). In addition, we do not find any evidence that available center-based alternatives are more effective than Head Start on average (e.g., Gormley et al., 2010; Barnett and Haskins, 2010). In the HSIS sample, around half of the control group children who enrolled in some other form of center-based care did so in either a state-funded prekindergarten program or a prekindergarten program based in the public schools.[11] While statistical power is limited, the null finding for Center Compliers suggests that concerns over Head Start's comparative effectiveness may be misplaced.

In addition to showing larger impacts in the Head Start year, we also find that the fade out in treatment effects over time is gradual, not rapid (Gibbs et al., 2011). This pattern closely resembles the observed fade out in other early childhood education studies (Magnuson et al., 2007; Leak et al., 2010). Further, while our estimates are imprecise, we find impacts between 0.10 to 0.15 for Home Complier children in first grade. These point estimates are very close to those in Deming (2009), who estimates Head Start impacts of 0.15 for children aged 5 to 6 and 0.13 for children aged 7 to 10.[12] Importantly, Deming (2009) observes outcomes for these same children in young adulthood, showing large long-term impacts. It is therefore possible that future follow up from the Head Start Impact Study will also find meaningful long-term impacts despite treatment effect fade out on short-term outcomes.

More generally, our analysis highlights the critical role that variation in counterfactual care type plays in early childhood education evaluations. Duncan and Magnuson (2013) argue that improving counterfactual conditions are a primary reason for a sharp decline in reported impacts of early childhood education interventions over the last half-century. We not only provide evidence consistent with this claim, but also outline a framework for re-analyzing other early childhood education studies to create comparable estimates. Of course, the issue of variation in counterfactual treatments is common in program evaluation settings, including for alternative schools (Bloom and Unterman, 2014) and job training programs (Heckman et al., 2000; Schochet et al., 2008). Our approach could easily be extended to these settings as well.

There are several promising avenues for future research. First, at present, we only analyze a single outcome of HSIS and analyze each follow-up year separately rather than jointly. Recent work

[11] Like Head Start, these publicly funded programs typically feature minimum standards for important structural aspects of program quality such as teacher preparation, teacher-child ratio and curricula. This result is also consistent with a recent study in Tulsa that found that Head Start and a publicly funded prekindergarten program led to comparable school readiness (Jenkins et al., 2014) and with the larger literature comparing quality for publicly funded versus private preschool programs (Kagan, 1991; Morris and Helburn, 2000).

[12] The outcome in Deming (2009) combines PPVT with the Peabody Individual Achievement Tests (PIAT) for math and reading.

from Mattei et al. (2013) suggests that looking at multiple outcomes—either across different test scores or over time—could greatly improve inference for principal causal effects (see also Jo and Muthén, 2001). In addition, repeated measures of the same outcome would likely make different assumptions about missingness more plausible (see, for example, Frumento et al., 2012). Second, while we conduct extensive sensitivity and robustness checks, inference with finite mixture models is notoriously difficult. In investigations for very simple mixture models, we have found that standard estimators can behave poorly when mixture components are not well separated (Day, 1969; Feller et al., 2016). More work is needed to assess whether these same concerns apply to the much richer models we consider here, although Griffin et al. (2008) have taken an important step in this direction. Overall, the finite sample properties of these methods are not fully understood, which is a serious caveat to the findings and approach reported here. That being said, the stability of the results to sensitivity checks, consistent patterns across subgroups, and alignment with Kline and Walters (2015) are all encouraging.

In the end, our results support the argument that further efforts to improve the early skill development of US children through the expansion of publicly-funded preschool programs should be targeted toward those who are currently not enrolling their children in center-based programs (for discussion, see Ludwig and Phillips, 2010; Bassok et al., 2013; Cascio and Schanzenbach, 2013). Nationwide, over 40 percent of eligible children are served by Head Start programs (Schmit et al., 2013). Although the availability of state and local prekindergarten has grown in recent years, many low-income children still spend their preschool years in home-based settings. In 2011, approximately 42 percent of three- and four-year-old children from low-income families enrolled in center-based prekindergarten compared to 59 percent of their non-low income peers (Burgess et al., 2014). Based on our results, shifting children from home-based care into formal care will likely lead to much larger effects than shifting children between preschool programs.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics* (113), 231–263.

Abadie, A., J. D. Angrist, and G. Imbens (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica 70*(1), 91–117.

Administration for Children and Families (2014). Head Start program facts, fiscal year 2013. Available at https://eclkc.ohs.acf.hhs.gov/hslc/data/factsheets/docs/hs-program-fact-sheet-2013.pdf.

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The Economic Journal 114*, C52–C83.

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

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Aronow, P. M. and A. Carnegie (2013). Beyond LATE: Estimation of the average treatment effect with an instrumental variable. *Political Analysis 21*, 492–506.

Bafumi, J. and A. E. Gelman (2006). Fitting multilevel models when predictors and group effects correlate.

Barnard, J., C. E. Frangakis, J. L. Hill, and D. B. Rubin (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association 98*(462), 299–323.

Barnett, W. S. (1995). Long-Term Effects of Early Childhood Programs on Cognitive and School Outcomes. *The future of children 5*(3), 25.

Barnett, W. S. (2011). Effectiveness of Early Educational Intervention. *Science 333*(6045), 975–978.

Barnett, W. S., M. E. Carolan, J. H. Squires, and K. C. Brown (2014). State of Preschool 2013: First Look. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Barnett, W. S. and R. Haskins (2010). *Investing in young children: New directions in federal preschool and early childhood policy.* Washington, DC: The Brookings Institute.

Bassok, D., M. Fitzpatrick, and S. Loeb (2013). Does state preschool crowd-out private provision? The impact of universal preschool on the childcare sector in Oklahoma and Georgia. NBER Working Paper 18605.

Bitler, M., J. Gelbach, and H. Hoynes (2003). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review 96*(4), 988–1012.

Bitler, M., H. Hoynes, and T. Domina (2014). Experimental Evidence on Distributional Effects of Head Start. Working Paper.

Bloom, H. S., S. Raudenbush, and M. Weiss (2014). Using Multi-site Evaluations to Study Variation in Effects of Program Assignment.

Bloom, H. S. and R. Unterman (2014). Can small high schools of choice improve educational prospects for disadvantaged students? *Journal of Policy Analysis and Management 33*(2), 290–319.

Bloom, H. S. and C. Weiland (2014). To what extent do the effects of Head Start on enrolled children vary across sites? Working Paper.

Bordes, L., S. Mottelet, and P. Vandekerkhove (2006). Semiparametric Estimation of a Two-Component Mixture Model. *The Annals of Statistics 34*(3), 1204–1232.

Burgess, K., N. Chien, T. Morrissey, and K. Swenson (2014). Trends in the use of early care and education, 1995-2011: Descriptive analysis of child care arrangements from national survey data. Report from the Office of the Assistant Secretary for Plannng and Evaluation, US Department of Health and Human Services.

Carneiro, P. and R. Ginja (2014). Long term impacts of compensatory preschool on health and behavior: Evidence from Head Start. *AEJ Applied Economics*.

Cascio, E. U. and D. W. Schanzenbach (2013). The Impacts of Expanding Access to High-Quality Preschool Education. *Brookings Papers on Economic Activity*, 127–192.

Coulson, A. J. (2013). Preschool's anvil chorus. Cato Institute.

Cox, D. R. and C. A. Donnelly (2011). *Principles of applied statistics*. Cambridge University Press.

Currie, J. and D. Thomas (1993). Does Head Start make a difference?

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika 56*(3), 463–474.

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

Ding, P., Z. Geng, W. Yan, and X.-H. Zhou (2011). Identifiability and Estimation of Causal Effects by Principal Stratification With Outcomes Truncated by Death. *Journal of the American Statistical Association 106*(496), 1578–1591.

Duncan, G. J. and K. Magnuson (2013). Investing in Preschool Programs. *Journal of Economic Perspectives 27*(2), 109–132.

Elango, S., J. L. García, J. J. Heckman, and A. Hojman (2015). Early childhood education. Technical report, National Bureau of Economic Research Working Paper #21766.

Feller, A. (2015). *Essays in Public Policy and Causal Inference*. Ph. D. thesis, Harvard University.

Feller, A., E. Greif, L. Miratrix, and N. Pillai (2016). Principal stratification in the Twilight Zone: Weakly separated components in finite mixture models.

Firpo, S. (2007). Efficient Semiparametric Estimation of Quantile Treatment Effects. *Econometrica 75*(1), 259–276.

Frangakis, C. E. and D. B. Rubin (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika 86*(2), 365–379.

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Frölich, M. and B. Melly (2013). Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics 31*(3), 346–357.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes*. Springer.

Frumento, P., F. Mealli, B. Pacini, and D. B. Rubin (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association 107*(498), 450–466.

Frumento, P., F. Mealli, B. Pacini, and D. B. Rubin (2016). The Fragility of Standard Inferential Approaches in Principal Stratification Models Relative to Direct Likelihood Approaches. *Statistical Analysis and Data Mining*.

Fryer, R. G. and S. D. Levitt (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics 86*(2), 447–464.

Gallop, R., D. S. Small, J. Y. Lin, M. R. Elliott, M. Joffe, and T. R. Ten Have (2009). Mediation analysis with principal stratification. *Statistics in Medicine 28*(7), 1108–1130.

Garces, E., D. Thomas, and J. Currie (2002). Longer-Term Effects of Head Start. *The American Economic Review 92*(4), 999–1012.

Gelber, A. and A. Isen (2013). Children's schooling and parents' behavior: Evidence from the Head Start Impact Study. *Journal of Public Economics 101*(C), 25–38.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis 1*(3), 515–534.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis.* CRC press.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics 2*(4), 1360–1383.

Gibbs, C., J. Ludwig, and D. L. Miller (2011). Does Head Start do any lasting good? In *The War on Poverty: A 50-Year Retrospective.*

Gormley, W. T. (2007). Early childhood care and education: Lessons and puzzles. *Journal of Policy Analysis and Management 26*(3), 633–671.

Gormley, W. T., D. Phillips, S. Adelstein, and C. Shaw (2010). Head start's comparative advantage: Myth or reality? *Policy Studies Journal 38*(3), 397–418.

Griffin, B. A., D. F. McCaffrey, and A. R. Morral (2008). An application of principal stratification to control for institutionalization at follow-up in studies of substance abuse treatment programs. *The Annals of Applied Statistics 2*(3), 1034–1055.

Hall, P. and X.-H. Zhou (2003). Nonparametric Estimation of Component Distributions in a Multivariate Mixture. *The Annals of Statistics 31*(1), 201–224.

Heckman, J., N. Hohmann, J. Smith, and M. Khoo (2000). Substitution and dropout bias in social experiments: A study of an influential social experiment. *Quarterly Journal of Economics*, 651–694.

Heckman, J. J. (2006). Skill Formation and the Economics of Investing in Disadvantaged Children. *Science 312*(5782), 1900–1902.

Hill, J., J. Waldfogel, and J. Brooks-Gunn (2002). Differential effects of high-quality child care. *Journal of Policy Analysis and Management 21*(4), 601–627.

Hirano, K., G. W. Imbens, D. B. Rubin, and X. H. Zhou (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics 1*(1), 69–88.

Hoffman, M. D. and A. Gelman (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research 15*(Apr), 1593–1623.

Hsu, J. Y. and D. S. Small (2014). Discussion on "Dynamic treatment regimes: technical challenges and applications". Working Paper.

Hunter, D. R., S. Wang, and T. P. Hettmansperger (2007). Inference for mixtures of symmetric distributions. *The Annals of Statistics 35*(1), 224–251.

Huston, A. C., Y. E. Chang, and L. Gennetian (2002). Family and individual predictors of child care use by low-income families in different policy contexts. *Early Childhood Research Quarterly 17*(4), 441–469.

Ibrahim, J. G. (1990). Incomplete Data in Generalized Linear Models. *Journal of the American Statistical Association 85*(411), 765–769.

Imai, K., G. King, and E. A. Stuart (2008). Misunderstandings between Experimentalists and Observationalists about Causal Inference. *Journal of the Royal Statistical Society. Series A. Statistics in Society 171*(2), 481–502.

Imbens, G. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Imbens, G. W. and D. B. Rubin (1997a). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics 25*(1), 305–327.

Imbens, G. W. and D. B. Rubin (1997b). Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *The Review of Economic Studies 64*(4), 555–574.

Jenkins, J. M., G. Farkas, G. J. Duncan, M. Burchinal, and D. L. Vandell (2014). Head start at ages 3 and 4 versus head start followed by state pre-k: Which is more effective? Working Paper.

Jin, H. and D. B. Rubin (2009). Public schools versus private schools: Causal inference with partial compliance. *Journal of Educational and Behavioral Statistics 34*(1), 24–45.

Jo, B. (2002). Estimation of Intervention Effects with Noncompliance: Alternative Model Specifications. *Journal of Educational and Behavioral Statistics 27*(4), 385–409.

Jo, B. and B. O. Muthén (2001). Modeling of intervention effects with noncompliance: A latent variable approach for randomized trials. *New developments and techniques in structural equation modeling*, 57–87.

Jo, B. and E. A. Stuart (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine 28*(23), 2857–2875.

Joffe, M. M., D. Small, and C.-Y. Hsu (2007). Defining and Estimating Intervention Effects for Groups that will Develop an Auxiliary Outcome. *Statistical Science 22*(1), 74–97.

Kagan, S. L. (1991). Examining profit and nonprofit child care: An odyssey of quality and auspices. *Journal of Social Issues 47*(2), 87–104.

Kline, P. and C. Walters (2015). Evaluating public programs with close substitutes: The case of Head Start. Working Paper.

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

Leak, J., G. J. Duncan, W. Li, K. A. Magnuson, H. Schindler, and H. Yoshikawa (2010). Is Timing Everything? How Early Childhood Education Program Impacts Vary by Starting Age, Program Duration and Time Since the End of the Program. Working Paper.

Lewandowski, D., D. Kurowicka, and H. Joe (2009). Journal of Multivariate Analysis. *Journal of Multivariate Analysis 100*(9), 1989–2001.

Ludwig, J. and D. 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.

Ludwig, J. and D. A. Phillips (2010). Leave no (young) child behind: prioritizing access in early childhood education. In Ron Haskins and W. Steven Barnett (Ed.), *Investing in Young Children: New Directions in Federal Preschool and Early Childhood Policy*. Brookings and NIEER.

Magnuson, K. A., C. Ruhm, and J. Waldfogel (2007). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly 22*(1), 18–38.

Mattei, A., F. Li, F. Mealli, et al. (2013). Exploiting multiple outcomes in bayesian principal stratification analysis with application to the evaluation of a job training program. *The Annals of Applied Statistics 7*(4), 2336–2360.

McCoy, D. C., M. C. Connors, P. A. Morris, H. Yoshikawa, and A. H. Friedman-Krauss (2014). Neighborhood economic disadvantage and childrens cognitive and social-emotional development: Exploring head start classroom quality as a mediating mechanism. Working Paper.

Mealli, F., G. W. Imbens, S. Ferro, and A. Biggeri (2004). Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics*.

Mealli, F. and B. Pacini (2008). Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness. *Computational Statistics & Data Analysis 53*, 507–516.

Mealli, F. and B. Pacini (2013). Using Secondary Outcomes to Sharpen Inference in Randomized Experiments With Noncompliance. *Journal of the American Statistical Association 108*(503), 1120–1131.

Miller, E. B., G. Farkas, D. L. Vandell, and G. J. Duncan (2014). Do the Effects of Head Start Vary by Parental Preacademic Stimulation? *Child Development 85*(4), 1385–1400.

Morris, J. R. and S. W. Helburn (2000). Child care center quality differences: The role of profit status, client preferences, and trust. *Nonprofit and Voluntary Sector Quarterly 29*(3), 377–399.

National Forum on Early Childhood Policy and Programs (2010). Understanding the head start impact study.

Neyman, J. (1923 [1990]). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472.

Page, L. C. (2012). Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness 5*(3), 215–244.

Page, L. C., A. Feller, T. Grindal, L. Miratrix, and M. A. Somers (2015). Principal Stratification: A Tool for Understanding Variation in Program Effects Across Endogenous Subgroups. *American Journal of Evaluation 36*(4), 514–531.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 71–110.

Puma, M., S. H. Bell, R. Cook, C. Heid, and G. Shapiro (2010a). Head Start Impact Study. Final Report. *HHS, Administration for Children & Families*.

Puma, M., S. H. Bell, R. Cook, C. Heid, and G. Shapiro (2010b). Head Start Impact Study. Technical Report. *HHS, Administration for Children & Families*.

Raudenbush, S. W. (2015). Estimation of Means and Covariance Components in Multi-site Randomized Trials.

Raudenbush, S. W., S. F. Reardon, and T. Nomi (2012). Statistical Analysis for Multisite Trials Using Instrumental Variables With Random Coefficients. *Journal of Research on Educational Effectiveness 5*(3), 303–332.

Reardon, S. F. and S. W. Raudenbush (2013). Under What Assumptions Do Site-by-Treatment Instruments Identify Average Causal Effects? *Sociological Methods & Research 42*(2), 143–163.

Romano, E., L. Babchishin, L. S. Pagani, and D. Kohen (2010). School readiness and later achievement: replication and extension using a nationwide canadian survey. *Developmental Psychology 46*(5), 995.

Rose, K. K. and J. Elicker (2010). Maternal child care preferences for infants, toddlers, and preschoolers: the disconnect between policy and preference in the USA. *Community, Work & Family 13*(2), 205–229.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Rubin, D. B. (1980). Comment on "randomization analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association 75*(371), 591–593.

Rubin, D. B. et al. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics 12*(4), 1151–1172.

Schmit, S., H. Matthews, S. Smith, and T. Robbins (2013). Investing in Young Children: A Fact Sheet on Early Care and Education Participation, Access, and Quality. Fact Sheet. New York, NY: National Center for Children in Poverty. Washington, DC: Center for Law and Social Policy.

Schochet, P., M. Puma, and J. Deke (2014). Understanding Variation in Treatment Effects in Education Impact Evaluations: An Overview of Quantitative Methods. (NCEE 20144017) Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

Schochet, P. Z. (2013). Student Mobility, Dosage, and Principal Stratification in School-Based RCTs. *Journal of Educational and Behavioral Statistics 38*(4), 323–354.

Schochet, P. Z. and J. Burghardt (2007). Using Propensity Scoring to Estimate Program-Related Subgroup Impacts in Experimental Program Evaluations. *Evaluation Review 31*(2), 95–120.

Schochet, P. Z., J. Burghardt, and S. McConnell (2008). Does Job Corps Work? Impact Findings from the National Job Corps Study. *The American Economic Review 98*(5), 1864–1886.

Scott-Clayton, J. and V. Minaya (2014). Should student employment be subsidized? conditional counterfactuals and the outcomes of work-study participation. National Bureau of Economic Research, Working Paper w20329.

Shager, H. M., H. S. Schindler, K. A. Magnuson, G. J. Duncan, H. Yoshikawa, and C. M. D. Hart (2013). Can Research Design Explain Variation in Head Start Research Results? A Meta-Analysis of Cognitive and Achievement Outcomes. *Educational Evaluation and Policy Analysis 35*(1), 76–95.

Stan Development Team (2014). Stan: A C++ library for probability and sampling, version 2.3.

Walters, C. R. (2015). Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start. *American Economic Journal: Applied Economics 7*(4), 76–102.

Westinghouse Learning Corporation (1969). *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development, Volume 1: Report to the Office of Economic Opportunity*. Athens, Ohio: Westinghouse Learning Corporation and Ohio University.

Whitehurst, G. J. (2013a). Can we be hard-headed about preschool? a look at head start. Brookings Institution.

Whitehurst, G. J. (2013b). Obama's preschool plan. Brookings Institution.

Zhai, F., J. Brooks-Gunn, and J. Waldfogel (2011). Head Start and urban children's school readiness: A birth cohort study in 18 cities. *Developmental Psychology 47*(1), 134–152.

Zhai, F., J. Brooks-Gunn, and J. Waldfogel (2014). Head Start's Impact Is Contingent on Alternative Type of Care in Comparison Group. *Developmental Psychology*.

Zhang, J. L. and D. B. Rubin (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics 28*(4), 353–368.

Zhang, J. L., D. B. Rubin, and F. Mealli (2009). Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification. *Journal of the American Statistical Association 104*(485), 166–176.

Zigler, E. and S. Muenchow (1992). *Head Start: The inside story of America's most successful educational experiment.* Basic Books.

Table 1: Covariate Balance at Baseline

	Control Mean	T-C Diff.	Norm. Diff.
Child Characteristics			
PPVT pre-test (std.)	0.03	-0.05	-0.04
Bottom third by pre-test	0.32	0.02	0.03
Three-year old	0.55	—	0.01
Male	0.51	—	-0.01
Black	0.30	0.01	0.02
Hispanic	0.37	0.01	0.01
Dual-Language Learner	0.29	0.01	0.03
Special needs	0.11	0.03	0.08
Caregiver and Family Characteristics			
Caregiver age: <25	0.32	-0.02	-0.05
Caregiver age: 25-29	0.31	—	—
Caregiver age: 30-39	0.29	0.01	0.02
Caregiver age: 40+	0.07	0.02	0.06
Teen mother	0.19	-0.03	-0.07
High school dropout	0.39	-0.02	-0.04
Only high school diploma/GED	0.33	0.01	0.02
Married	0.45	-0.01	-0.01
Previously married	0.16	—	—
Urban	0.84	—	—
Family risk: medium/high	0.22	0.03	0.06
Lives with both biological parents	0.49	—	—
Recent immigrant	0.19	—	0.01
Any older sibling attended Head Start	0.37	0.04	0.09
Oldest child	0.45	-0.03	-0.06
Head Start Center of Random Assignment Characteristics			
Provides transportation	0.63	—	—
At least four home visits per year	0.21	—	-0.01
Full day child care	0.64	—	0.01
Student-teacher ratio	6.75	-0.02	-0.01
All teachers certified in early childhood	0.41	—	—
All teachers have mentors	0.46	—	—
Center is always filled	0.48	—	—
Number of children randomized	17	—	—
Neighborhood and State Characteristics			
Percent in poverty	0.25	—	—
Percent minority	0.44	—	—
Percent unemployed	0.11	—	—
Percent commute by car	0.82	—	—
Number of crimes per 1000 people	44	0.1	0.01
State has DOE Pre-K	0.64	—	0.01
State per-child spending ($'000)	3.9	—	0.01
State Head Start teacher salary ($'000)	21.8	—	0.01

Notes: Section 3.3 discusses the Normalized Difference. For clarity, 0.00 is denoted by '—'.

Table 2: Child care setting by treatment group

	Treatment	Control	Difference
Head Start	0.77	0.11	0.66
Other center-based care	0.08	0.26	-0.18
Home-based care	0.09	0.47	-0.38
Missing	0.06	0.16	-0.10
Head Start (admin.)	0.81	0.12	0.69

Notes: Child care setting is based on responses from the Spring 2003 parent reports. “Head Start (admin.)” refers to the administrative records collected as part of HSIS and is the compliance rate used in Puma et al. (2010a).

Table 3: Possible principal strata in the Head Start Impact Study

		$Z = 0$	
		Head Start	**Not Head Start**
$Z = 1$	**Head Start**	Always Head Start	Complier
	Not Head Start	*(Defier)*	Never Head Start

(a) **Binary D^*: Head Start vs. No Head Start.**

		$Z = 0$		
		Head Start	**Center Care**	**Home Care**
$Z = 1$	**Head Start**	Always Head Start	Center Complier	Home Complier
	Center Care	*(A)*	Always Center Care	*(B)*
	Home Care	*(C)*	*(D)*	Always Home Care

(b) **Multi-valued D: Head Start, Other Center-based care, Home-based care.**

Table 4: Distribution of Principal Strata

Noncompliers			**Compliers**	
Always HS	Always Center	Always Home	Center Complier	Home Complier
0.11	0.11	0.12	0.20	0.45

Notes: Posterior medians, with missing care type imputed. See Appendix C.2.

Table 5: Relationship between Observed Care Type and Principal Strata

Z	D^*	Possible Principal Strata
1	HS	Always Head Start, Complier (treat)
1	Not HS	Never Head Start
0	HS	Always Head Start
0	Not HS	Never Head Start, Complier (control)

(a) Binary D^*: Head Start vs. No Head Start.

Z	D	Possible Principal Strata
1	HS	Always Head Start, Center Complier (treat), Home Complier (treat)
1	Center	Always Center
1	Home	Always Home
0	HS	Always Head Start
0	Center	Always Center, Center Complier (control)
0	Home	Always Home, Home Complier (control)

(b) Multi-valued D: Head Start, Other Center-based care, Home-based care.

Table 6: Covariate Means by Principal Stratum

	Always	Always	Always	Center	Home
	Head Start	Center	Home	Complier	Complier
Child Characteristics					
PPVT pre-test (std.)	-0.20	0.24	-0.03	0.12	-0.05
Bottom third by pre-test	0.36	0.28	0.35	0.30	0.35
Three-year old	0.63	0.43	0.53	0.47	0.59
Male	0.54	0.53	0.57	0.48	0.48
Black	0.35	0.40	0.23	0.31	0.30
Hispanic	0.42	0.33	0.39	0.35	0.37
Dual-Language Learner	0.37	0.28	0.27	0.33	0.29
Special needs	0.15	0.17	0.12	0.13	0.11
Caregiver and Family Characteristics					
Caregiver age: <25	0.29	0.31	0.38	0.24	0.32
Caregiver age: 25-29	0.29	0.31	0.30	0.35	0.31
Caregiver age: 30-39	0.34	0.28	0.26	0.30	0.29
Caregiver age: 40+	0.08	0.09	0.07	0.11	0.08
Teen mother	0.16	0.19	0.20	0.15	0.17
High school dropout	0.44	0.27	0.47	0.35	0.38
Only high school diploma/GED	0.28	0.35	0.31	0.30	0.36
Married	0.46	0.42	0.47	0.45	0.44
Previously married	0.14	0.18	0.16	0.17	0.16
Urban	0.90	0.87	0.86	0.86	0.81
Family risk: medium/high	0.27	0.16	0.22	0.24	0.25
Lives with both biological parents	0.51	0.47	0.48	0.48	0.51
Recent immigrant	0.23	0.20	0.18	0.22	0.17
Any older sibling attended Head Start	0.40	0.34	0.38	0.34	0.43
Oldest child	0.43	0.47	0.45	0.50	0.39
Head Start Center of Random Assignment Characteristics					
Provides transportation	0.44	0.61	0.60	0.62	0.68
At least four home visits per year	0.15	0.17	0.21	0.18	0.25
Full day child care	0.69	0.75	0.59	0.67	0.61
Student-teacher ratio	6.66	6.89	6.58	7.07	6.64
All teachers certified in early childhood	0.50	0.43	0.41	0.44	0.38
All teachers have mentors	0.38	0.49	0.43	0.46	0.48
Center is always filled	0.50	0.43	0.44	0.48	0.49
Number of children randomized	14	18	15	16	18
Neighborhood and State Characteristics					
Percent in poverty	0.27	0.25	0.23	0.27	0.24
Percent minority	0.55	0.49	0.40	0.45	0.40
Percent unemployed	0.12	0.11	0.10	0.11	0.10
Percent commute by car	0.72	0.77	0.82	0.81	0.85
Number of crimes per 1000 people	49	45	42	47	43
State has DOE Pre-K	0.72	0.69	0.59	0.68	0.62
State per-child spending ($'000)	3.4	3.8	3.4	4.1	4.2
State Head Start teacher salary ($'000)	21.1	21.7	21.3	21.9	22.1

Notes: Covariate means based on multiply imputed stratum membership.

Table 7: Impacts in the Head Start Year

A. ITT Model	
ITT	**0.14** **(0.11, 0.16)**
B. IV Model	
Overall $LATE$	**0.18** **(0.14, 0.23)**
C. Principal Stratification Model	
$LATE$ for Center Compliers	0.00 (-0.13, 0.14)
$LATE$ for Home Compliers	**0.23** **(0.15, 0.30)**
$\mathbb{P}\{LATE_{\text{hc}} > LATE_{\text{cc}}\}$	0.99

Notes: Point estimates are posterior medians, with 2.5 and 97.5 quantiles of posterior distribution in parentheses. 95% posterior intervals that exclude zero are printed in bold. See Appendix B for estimation details. All treatment effects, including $LATE_{\text{cc}}$ and $LATE_{\text{hc}}$, are allowed to vary by Head Start center of random assignment. Note that all three models are estimated separately. The estimates from the principal stratification model imply $LATE = 0.16$ (0.12, 0.21) and $ITT = 0.11$ (0.08, 0.13), which are slightly lower than the estimates from the separate models.

Table 8: Impacts in the Head Start Year for Select Subgroups

	Center Compliers	Home Compliers
Panel A. Bottom Third on Pre-Test		
Bottom Third	0.19	**0.30**
	(-0.09, 0.47)	**(0.16, 0.45)**
Not Bottom Third	-0.06	**0.21**
	(-0.24, 0.16)	**(0.08, 0.31)**
Panel B. DLL Status		
DLL Students	0.06	**0.36**
	(-0.33, 0.42)	**(0.23, 0.49)**
Non-DLL Students	-0.04	**0.15**
	(-0.20, 0.12)	**(0.08, 0.23)**

Notes: Point estimates are posterior medians, with 2.5 and 97.5 quantiles of posterior distribution in parentheses. 95% posterior intervals that exclude zero are printed in bold. Estimates are shown in effect size units, so point estimates might not average to the pooled estimate due to different outcome standard deviations. See Appendix B for estimation details. All treatment effects are allowed to vary by Head Start center of random assignment.

Table 9: Sensitivity Analysis for Impacts in the Head Start Year

	Normal; Complete case	Student t_7; All observations
$LATE$ for Center Compliers	0.03 (-0.07, 0.15)	0.04 (0.08)
$LATE$ for Home Compliers	**0.21** **(0.15, 0.27)**	**0.21** **(0.15, 0.26)**
$\mathbb{P}\{LATE_{\text{hc}} > LATE_{\text{cc}}\}$	0.98	0.96

Notes: Point estimates are posterior medians, with 2.5 and 97.5 quantiles of posterior distribution in parentheses. 95% posterior intervals that exclude zero are printed in bold. See Appendix B for estimation details. All treatment effects, including $LATE_{\text{cc}}$ and $LATE_{\text{hc}}$, are allowed to vary by Head Start center of random assignment.

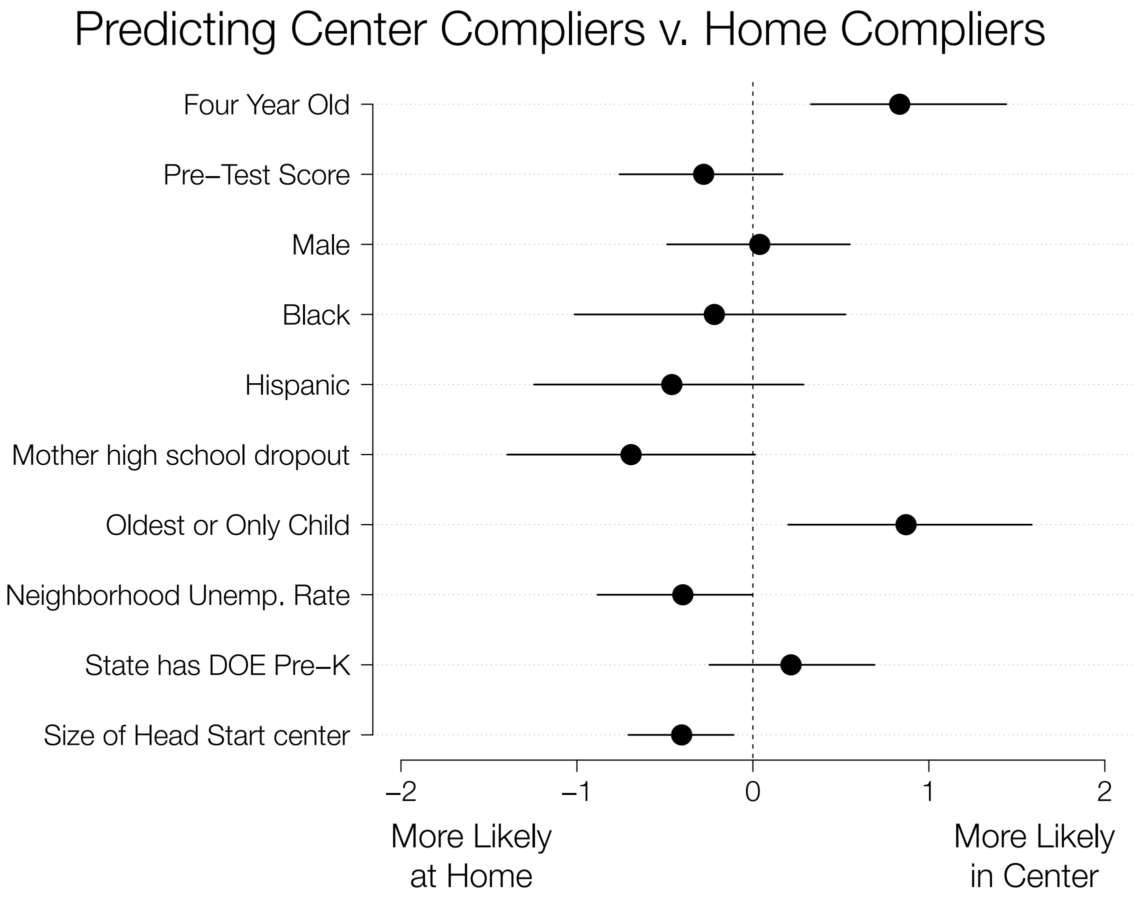


Figure 1: Logistic regression coefficients predicting Center vs. Home Compliers, generated from a multinomial logistic regression predicting all types. All continuous covariates are standardized. Point estimates and error bars show posterior medians and 95% credible intervals.

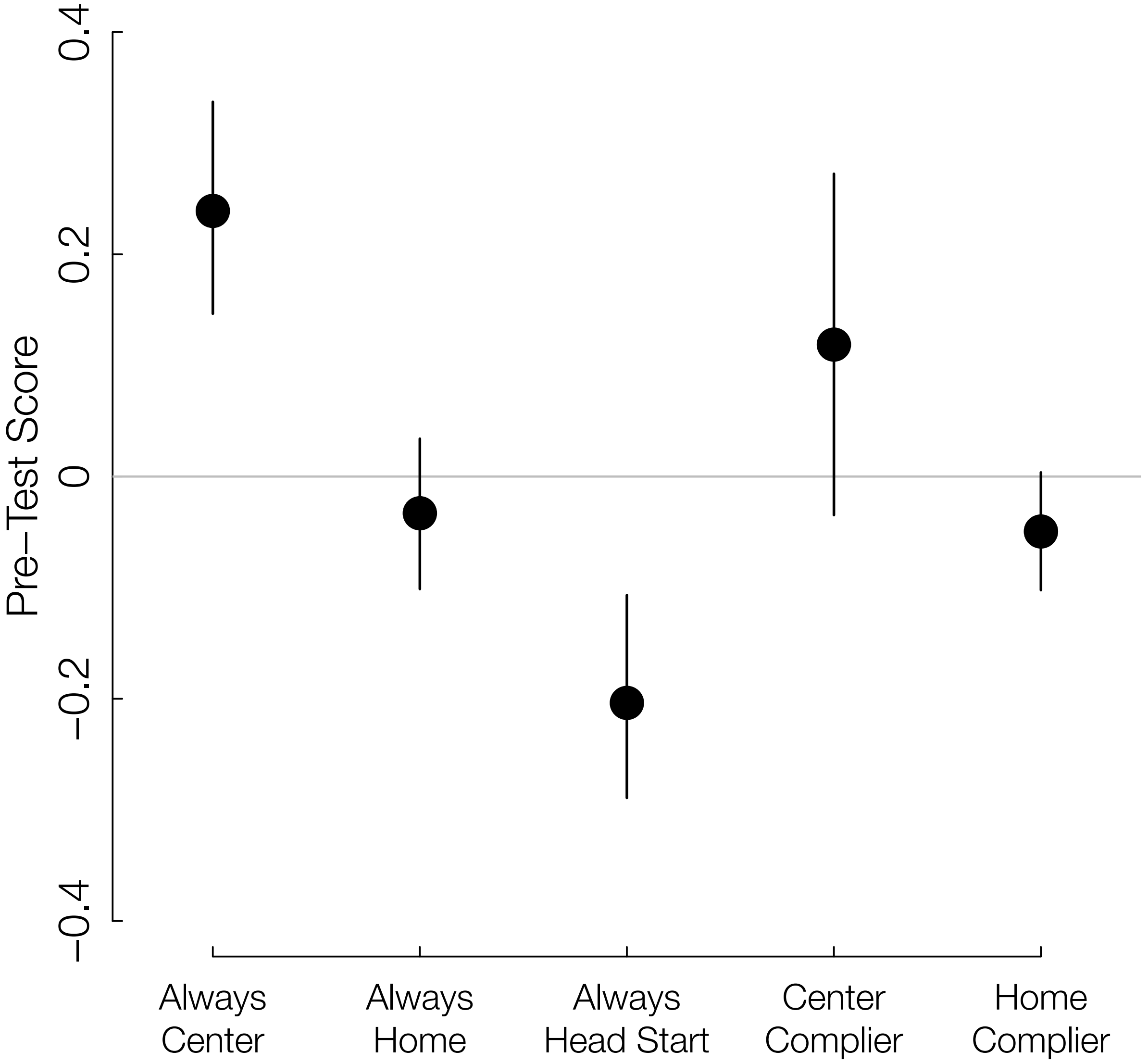
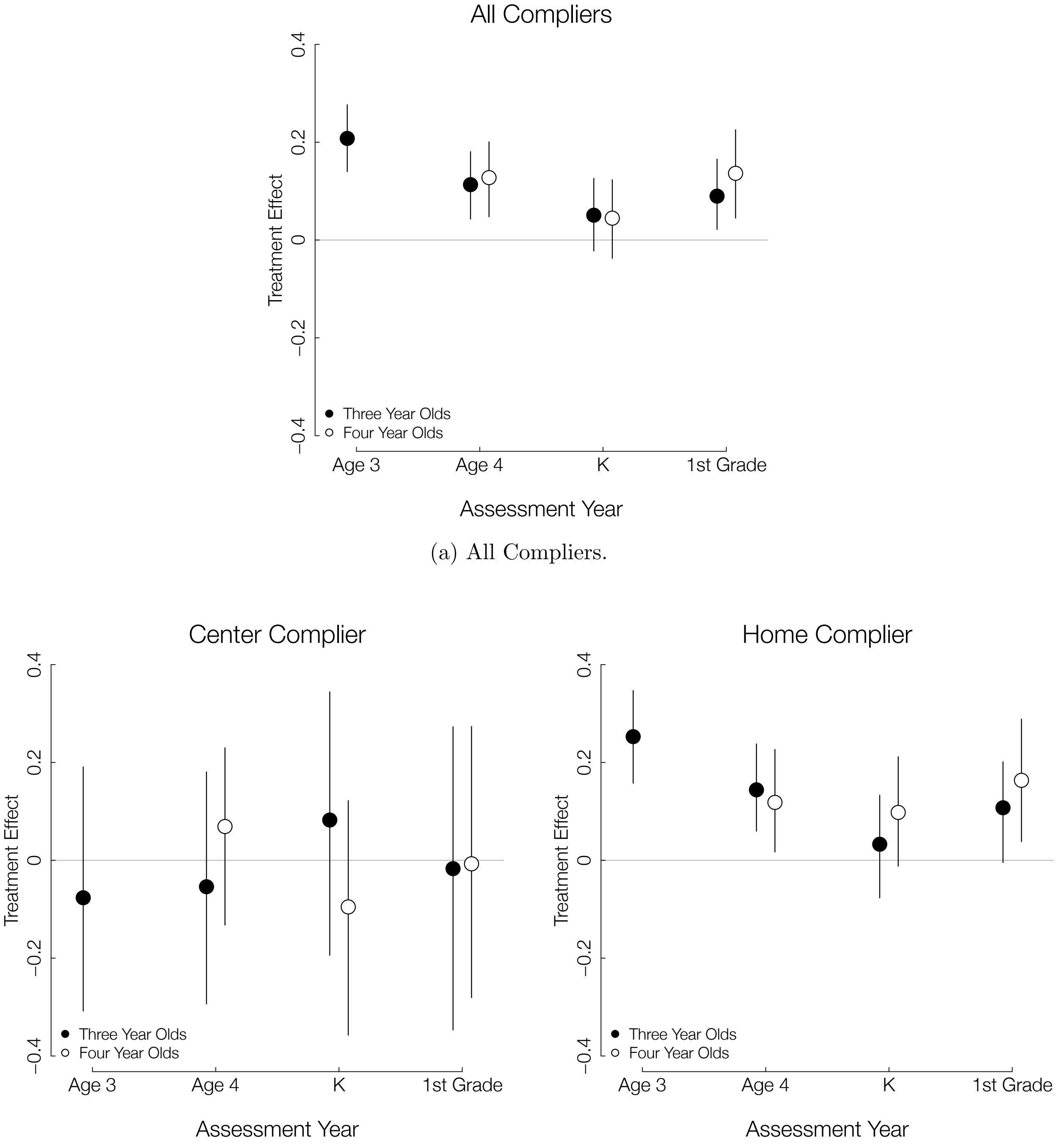


Figure 2: PPVT pre-test score by principal stratum. Point estimates and error bars show posterior medians and 95% credible intervals.

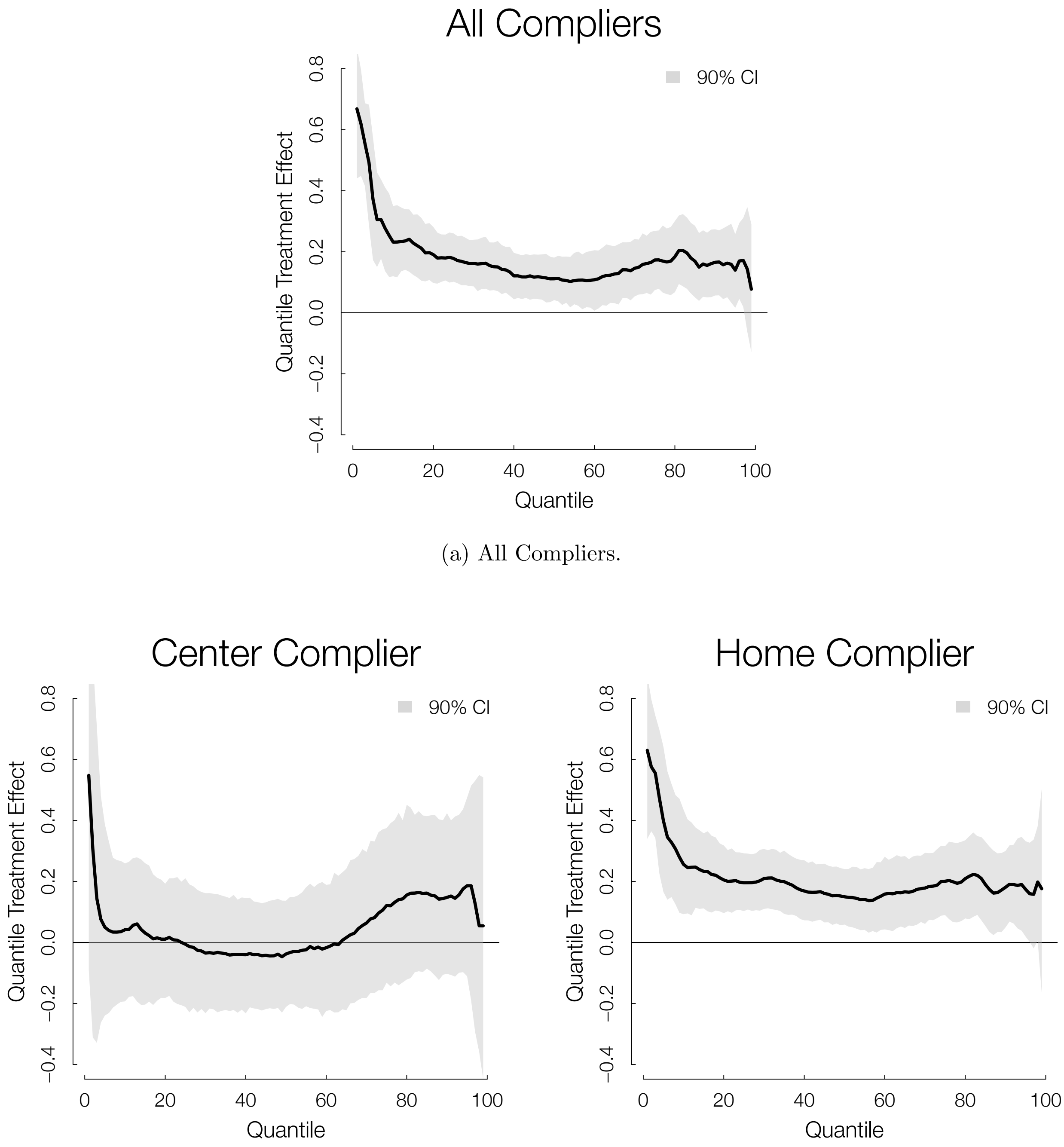


(a) All Compliers.

(b) Center Compliers.

(c) Home Compliers.

Figure 3: Impact estimates on PPVT by principal stratum and by three- and four-year-old cohort for each assessment year. Point estimates and error bars show posterior medians and 95% credible intervals. Effect sizes are calculated separately for each cohort in each assessment year.



(a) All Compliers.

(b) Center Compliers.

(c) Home Compliers.

Figure 4: Quantile treatment estimates on PPVT by principal stratum for the Head Start year, with approximate 90 percent credible intervals.

Online Appendix

A Identification for Outcome Distributions

A.1 Instrumental Variables

We begin with the IV case, following Imbens and Rubin (1997b) and Abadie (2003). Formally, let $g_{sz}(y)$ be the distribution of $Y_i(z)$ for principal stratum s. Table 3a shows the possible strata for binary D^* under the monotonicity assumption (Assumption IV-1): Always Head Start, Never Head Start, and Compliers. Under the exclusion restrictions (Assumption IV-2), the following outcome distributions are equal: $g_{\text{ahs}\,0}(y) = g_{\text{ahs}\,1}(y)$ and $g_{\text{nhs}\,0}(y) = g_{\text{nhs}\,1}(y)$. This yields four possible stratum outcome distributions: $g_{\text{ahs}}(y)$, $g_{\text{nhs}}(y)$, $g_{\text{co}\,0}(y)$, and $g_{\text{co}\,1}(y)$. Finally, let $f_{zd}(y)$ be the distribution of Y_i^{obs} in the subsample defined by $Z_i = z$ and $D_i^{*,obs} = d$. Table 5a shows the relationship between the observed and latent distributions. We can write these relationships mathematically:

$$\begin{aligned}
f_{11}(y) &= \frac{\pi_{\text{co}}}{\pi_{\text{co}} + \pi_{\text{ahs}}} g_{\text{co}\,1}(y) + \frac{\pi_{\text{ahs}}}{\pi_{\text{co}} + \pi_{\text{ahs}}} g_{\text{ahs}}(y) \\
f_{10}(y) &= g_{\text{nhs}}(y) \\
f_{01}(y) &= g_{\text{ahs}}(y) \\
f_{00}(y) &= \frac{\pi_{\text{co}}}{\pi_{\text{co}} + \pi_{\text{nhs}}} g_{\text{co}\,0}(y) + \frac{\pi_{\text{nhs}}}{\pi_{\text{co}} + \pi_{\text{nhs}}} g_{\text{nhs}}(y)
\end{aligned}$$

Simply re-arranging terms yields:

$$\begin{aligned}
g_{\text{ahs}}(y) &= f_{01}(y) \\
g_{\text{nhs}}(y) &= f_{10}(y) \\
g_{\text{co}\,0}(y) &= \frac{\pi_{\text{co}} + \pi_{\text{nhs}}}{\pi_{\text{co}}} f_{00}(y) - \frac{\pi_{\text{nhs}}}{\pi_{\text{co}}} g_{\text{nhs}}(y) \\
g_{\text{co}\,1}(y) &= \frac{\pi_{\text{co}} + \pi_{\text{ahs}}}{\pi_{\text{co}}} f_{11}(y) - \frac{\pi_{\text{ahs}}}{\pi_{\text{co}}} g_{\text{ahs}}(y)
\end{aligned}$$

Since we can non-parametrically identify π_s for all s in $\{\text{ahs}, \text{nhs}, \text{co}\}$ and $f_{zd}(y)$ for $Z_i \in \{0, 1\}$ and $D^*i \in \{0, 1\}$, we can therefore non-parametrically identify the relevant $g_{sz}(y)$.

A.2 Principal Strata

We now extend this logic to the full principal stratification problem. Table 3b shows the five possible principal strata under Assumptions PS-1a and PS-1b. Under the exclusion restrictions (Assumption PS-2), we have the following equalities: $g_{\text{ahs}\,0}(y) = g_{\text{ahs}\,1}(y)$, $g_{\text{ac}\,0}(y) = g_{\text{ac}\,1}(y)$, and $g_{\text{ah}\,0}(y) = g_{\text{ah}\,1}(y)$. Table 5b shows the relationship between the observed and latent distributions with multi-valued D. We again describe these mathematically:

$$\begin{aligned}
f_{1\,\text{HS}}(y) &= \frac{\pi_{\text{ahs}}}{\pi_{\text{ahs}}+\pi_{\text{cc}}+\pi_{\text{hc}}}\ g_{\text{ahs}}(y) + \frac{\pi_{\text{cc}}}{\pi_{\text{ahs}}+\pi_{\text{cc}}+\pi_{\text{hc}}}\ g_{\text{cc}\,1}(y) + \frac{\pi_{\text{hc}}}{\pi_{\text{ahs}}+\pi_{\text{cc}}+\pi_{\text{hc}}}\ g_{\text{hc}\,1}(y)\\
f_{1\,\text{Center}}(y) &= g_{\text{ac}}(y)\\
f_{1\,\text{Home}}(y) &= g_{\text{ah}}(y)\\
f_{0\,\text{HS}}(y) &= g_{\text{ahs}}(y)\\
f_{0\,\text{Center}}(y) &= \frac{\pi_{\text{ac}}}{\pi_{\text{ac}}+\pi_{\text{cc}}}\ g_{\text{ac}}(y) + \frac{\pi_{\text{cc}}}{\pi_{\text{ac}}+\pi_{\text{cc}}}\ g_{\text{cc}\,0}(y)\\
f_{0\,\text{Home}}(y) &= \frac{\pi_{\text{ah}}}{\pi_{\text{ah}}+\pi_{\text{hc}}}\ g_{\text{ah}}(y) + \frac{\pi_{\text{hc}}}{\pi_{\text{ah}}+\pi_{\text{hc}}}\ g_{\text{hc}\,0}(y)
\end{aligned}$$

Re-arranging terms yields five of the seven needed distributions:

$$\begin{aligned}
g_{\text{ac}}(y) &= f_{1\,\text{Center}}(y)\\
g_{\text{ah}}(y) &= f_{1\,\text{Home}}(y)\\
g_{\text{ahs}}(y) &= f_{0\,\text{HS}}(y)\\
g_{\text{cc}\,0}(y) &= \frac{\pi_{\text{ac}}+\pi_{\text{cc}}}{\pi_{\text{cc}}} f_{0\,\text{Center}}(y) - \frac{\pi_{\text{ac}}}{\pi_{\text{cc}}} g_{\text{ac}}(y)\\
g_{\text{hc}\,0}(y) &= \frac{\pi_{\text{ah}}+\pi_{\text{hc}}}{\pi_{\text{hc}}} f_{0\,\text{Home}}(y) - \frac{\pi_{\text{ah}}}{\pi_{\text{hc}}} g_{\text{ah}}(y)
\end{aligned}$$

Following the same logic as in the IV case, these outcome distributions are non-parametrically identified. However, we still need the outcome distributions for Compliers under treatment, $g_{\text{cc}\,1}(y)$ and $g_{\text{hc}\,1}(y)$. Since we observe $g_{\text{ahs}}(y)$ in the control group, we can isolate these outcome distributions by further "backing out" $g_{\text{ahs}}(y)$ from the three-component mixture of $f_{1\,\text{HS}}(y)$:

$$f^*_{1\,\text{HS}}(y) = \frac{\pi_{\text{ahs}}+\pi_{\text{cc}}+\pi_{\text{hc}}}{\pi_{\text{cc}}+\pi_{\text{hc}}} f_{1\,\text{HS}}(y) - \frac{\pi_{\text{ahs}}}{\pi_{\text{cc}}+\pi_{\text{hc}}} g_{\text{ahs}}(y).$$

We are then left with a classic two-component finite mixture model for $f^*_{1\,\text{HS}}(y)$:

$$f^*_{1\,\text{HS}}(y) = \phi\ g_{\text{hc}\,1}(y) + (1-\phi)\ g_{\text{cc}\,1}(y),$$

with known mixing proportion $\phi = \frac{\pi_{\text{hc}}}{\pi_{\text{cc}}+\pi_{\text{hc}}}$. In general, the component densities in a two-component finite mixture model are not identifiable without additional restrictions (see, e.g., Hall and Zhou, 2003). See the main text for additional discussion.

B Estimation

B.1 Prior distributions

Before computation, the outcome was centered by the global mean and re-scaled by the standard deviation of the observed outcomes for the control group, so that all units were in terms of "effect size." All covariates are either on a binary/unit scale or have been standardized. Throughout we use default, weakly informative priors on all model parameters (Gelman et al., 2008). For regression coefficients, all main coefficients have independent $N(0, 1.5^2)$ priors; stratum-by-covariate interaction terms have a tighter $N(0, 0.25^2)$ prior. For multinomial logistic regressions, all coefficients have Cauchy(0, 2.5) priors (Gelman et al., 2008). All standard deviations have half-Cauchy priors, with the scale set to 1, unless otherwise noted (Gelman, 2006). This effectively places a flat prior over the space of 0 to 2, while allowing for much larger standard deviations if there is a strong signal in the data. As the outcome has already been standardized to have a global standard deviation of 1 (for the control group), an observed model standard deviation greater than 2 would suggest an especially bad model fit. The prior for the site-level standard deviation is also given a half-Cauchy(0,1) distribution. Following the recommendation among Stan developers, all random effect correlation matrices are given weak LKJ priors (Lewandowski et al., 2009), which have more attractive properties than the more standard inverse-Wishart distribution and which slightly favor an identity correlation matrix (Stan Development Team, 2014). Finally, missing pre-tests and outcomes have a $N(0, 1)$ prior, on the same order as the standardized outcome distribution.

B.2 ITT

For the ITT model, we use a standard varying intercept/varying slope model (i.e., a "random effects" model), which accounts for center-level variation via multilevel modeling. Following Bloom et al. (2014), we also estimate separate residual variances for the treatment and control groups. Both the intercept, α_j, and the treatment effect, τ_j, vary by site.

B.3 Estimating LATE

Imbens and Rubin (1997a) proposed a model-based estimation strategy for instrumental variables models as an alternative to the standard Wald/Two-Stage Least Squares estimator. See Imbens and Rubin (2015) for a textbook discussion of this approach. The key idea is that the usual ratio estimators ignore individual-level information about compliance status, since they are based solely on sample averages.

Incorporating this information requires the use of a full likelihood:

$$\begin{aligned}\mathcal{L}_{obs}(\theta \mid \mathbf{y}, \mathbf{x}, \mathbf{d}, \mathbf{z}) \quad = \quad & \prod_{i:D_i=0,Z_i=1} \pi_{i:\text{nt}}\ g_{\text{nt}}(y_i \mid \theta, \mathbf{x}_i) \times \\ & \prod_{i:D_i=1,Z_i=0} \pi_{i:\text{at}}\ g_{\text{at}}(y_i \mid \theta, \mathbf{x}_i) \times \\ & \prod_{i:D_i=0,Z_i=0} \Big\{\pi_{i:\text{nt}}\ g_{\text{nt}}(y_i \mid \theta, \mathbf{x}_i) + \\ & \qquad \pi_{i:\text{co}}\ g_{\text{co0}}(y_i \mid \theta, \mathbf{x}_i)\Big\} \times \\ & \prod_{i:D_i=1,Z_i=1} \Big\{\pi_{i:\text{at}}\ g_{\text{at}}(y_i \mid \theta, \mathbf{x}_i) + \\ & \qquad \pi_{i:\text{co}}\ g_{\text{co1}}(y_i \mid \theta, \mathbf{x}_i)\Big\}\end{aligned}$$

In our setting, we incorporate weak prior information and estimate these parameters in a Bayesian framework. Computationally, we jointly estimate two sub-models. The first is a multinomial logistic regression predicting principal stratum membership as a function of covariates:

$$\pi_{s|\mathbf{x}} \equiv \mathbb{P}(S_i = s \mid \theta, \mathbf{x}_i) = \frac{\exp(\gamma_s + \delta_s' \mathbf{x}_i)}{\sum_{s=1}^{K} \exp(\gamma_s + \delta_s' \mathbf{x}_i)}.$$

While we do not explore additional models here, we note that we could replace the multinomial logit model with other discrete choice models, such as a multinomial probit model.

The second is an outcome model, effectively a separate regression for each compliance type:

$$\begin{aligned} y_i \mid (S_i^* = \text{nt}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{nt}} + \beta_{\text{nt}}' \mathbf{x}_i, \sigma_{\text{nt}}^2\right) \\ y_i \mid (S_i^* = \text{at}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{at}} + \beta_{\text{at}}' \mathbf{x}_i, \sigma_{\text{at}}^2\right) \\ y_i \mid (S_i^* = \text{co}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{co}} + \beta_{\text{co}}' \mathbf{x}_i + \tau z_i, \sigma_{\text{co},z}^2\right) \end{aligned}$$

where we partially pool the coefficients, $\beta_{s,k} \sim N\left(\mu_{\beta,k}, \eta_k^2\right)$, for $k = 1, \ldots, K$. The variance term differs by principal stratum and, among Compliers, by treatment assignment (Bloom et al., 2014).

We then make two modifications to extend model-based IV to a multi-level setting. First, we estimate a varying-intercept/varying-slope model separately for each principal stratum, where $j[i]$

indicates the site j corresponding to child i:

$$\begin{aligned}
y_i \mid (S_i^* = \text{nt}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{nt}} + \beta_{\text{nt}}\mathbf{x}_i + \psi_{j[i]}, \sigma^2_{\text{nt}}\right) \\
y_i \mid (S_i^* = \text{at}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{at}} + \beta_{\text{at}}\mathbf{x}_i + \psi_{j[i]}, \sigma^2_{\text{at}}\right) \\
y_i \mid (S_i^* = \text{co}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{co}} + \beta_{\text{co}}\mathbf{x}_i + \psi_{j[i]} + \tau z_i + \omega_{j[i]} z_i, \sigma^2_{\text{co},z}\right)
\end{aligned}$$

$$\begin{pmatrix} \psi_j \\ \omega_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta^{ctr}\mathbf{w}_j \\ 0 \end{pmatrix}, \begin{pmatrix} \eta^2_\psi & \rho\eta_\psi\eta_\omega \\ \rho\eta_\psi\eta_\omega & \eta^2_\omega \end{pmatrix}\right)$$

Second, we adjust the multinomial logistic regression to include a site-specific intercept. This simple varying-intercept model is repeated across all three principal strata:

$$\mathbb{P}(S_i = s \mid \theta, \mathbf{x}_i) = \frac{\exp(\gamma_{s,j[i]} + \delta_s' \mathbf{x}_i)}{\sum_{s=1}^K \exp(\gamma_{s,j[i]} + \delta_s' \mathbf{x}_i)}$$

$$\begin{pmatrix} \gamma_{n,j} \\ \gamma_{a,j} \\ \gamma_{c,j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\gamma,n} + \delta_n^{ctr}\mathbf{w}_j \\ \mu_{\gamma,a} + \delta_a^{ctr}\mathbf{w}_j \\ \mu_{\gamma,c} + \delta_c^{ctr}\mathbf{w}_j \end{pmatrix}, \begin{pmatrix} \eta^2_{\gamma,n} & 0 & 0 \\ 0 & \eta^2_{\gamma,a} & 0 \\ 0 & 0 & \eta^2_{\gamma,c} \end{pmatrix}\right)$$

where the site-level coefficients, δ_s^{ctr}, vary across strata, as in the non-hierarchical model.

B.4 Estimating the Principal Stratification Model

To estimate the full principal stratification model, we simply expand the likelihood from the instrumental variable model to account for the additional latent groups:

$$
\begin{aligned}
\mathcal{L}_{obs}(\theta \mid \mathbf{y}, \mathbf{x}, \mathbf{d}, \mathbf{z}) \quad = \quad & \prod_{i:D_i=HS, Z_i=0} \Big\{ \pi_{i:\text{ahs}} \; g_{\text{ahs}}(y_i \mid \theta, \mathbf{x}_i) \Big\} \times \\
& \prod_{i:D_i=Center, Z_i=0} \Big\{ \pi_{i:\text{cc}} \; g_{\text{cc}\,0}(y_i \mid \theta, \mathbf{x}_i) + \pi_{i:\text{ac}} \; g_{\text{ac}}(y_i \mid \theta, \mathbf{x}_i) \Big\} \times \\
& \prod_{i:D_i=Home, Z_i=0} \Big\{ \pi_{i:\text{hc}} \; g_{\text{hc}\,0}(y_i \mid \theta, \mathbf{x}_i) + \pi_{i:\text{ah}} \; g_{\text{ah}}(y_i \mid \theta, \mathbf{x}_i) \Big\} \times \\
& \prod_{i:D_i=?, Z_i=0} \Big\{ \pi_{i:\text{cc}} \; g_{\text{cc}\,0}(y_i \mid \theta, \mathbf{x}_i) + \pi_{i:\text{hc}} \; g_{\text{hc}\,0}(y_i \mid \theta, \mathbf{x}_i) + \pi_{i:\text{ah}} \; g_{\text{ah}}(y_i \mid \theta, \mathbf{x}_i) \Big\} \times \\
& \prod_{i:D_i=HS, Z_i=1} \Big\{ \pi_{i:\text{ahs}} \; g_{\text{ahs}}(y_i \mid \theta, \mathbf{x}_i) + \pi_{i:\text{cc}} \; g_{\text{cc}\,1}(y_i \mid \theta, \mathbf{x}_i) + \pi_{i:\text{hc}} \; g_{\text{hc}\,1}(y_i \mid \theta, \mathbf{x}_i) \Big\} \times \\
& \prod_{i:D_i=Center, Z_i=1} \Big\{ \pi_{i:\text{ac}} \; g_{\text{ac}}(y_i \mid \theta, \mathbf{x}_i) \Big\} \times \\
& \prod_{i:D_i=Home, Z_i=1} \Big\{ \pi_{i:\text{ah}} \; g_{\text{ah}}(y_i \mid \theta, \mathbf{x}_i) \Big\} \times \\
& \prod_{i:D_i=?, Z_i=1} \Big\{ \pi_{i:\text{ac}} \; g_{\text{ac}}(y_i \mid \theta, \mathbf{x}_i) + \pi_{i:\text{ah}} \; g_{\text{ah}}(y_i \mid \theta, \mathbf{x}_i) \Big\}
\end{aligned}
$$

Table A4 gives this mapping in words. Section C describes the assumptions for an ignorable missingness mechanism. As discussed in the main text, the corresponding outcome models are:

$$
\begin{aligned}
y_i^{\text{obs}} \mid (S_i = \text{ahs}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{ahs}} + \beta_{\text{ahs}}\mathbf{x}_i + \psi_{j[i]}, \sigma^2_{\text{ahs}}\right) \\
y_i^{\text{obs}} \mid (S_i = \text{ac}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{ac}} + \beta_{\text{ac}}\mathbf{x}_i + \psi_{j[i]}, \sigma^2_{\text{ac}}\right) \\
y_i^{\text{obs}} \mid (S_i = \text{ah}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{ah}} + \beta_{\text{ah}}\mathbf{x}_i + \psi_{j[i]}, \sigma^2_{\text{ah}}\right) \\
y_i^{\text{obs}} \mid (S_i = \text{cc}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{cc}} + \beta_{\text{cc}}\mathbf{x}_i + \psi_{j[i]} + \tau_{\text{cc}} z_i + \omega_{j[i],\text{cc}} z_i, \sigma^2_{\text{cc},z}\right) \\
y_i^{\text{obs}} \mid (S_i = \text{hc}, \theta, \mathbf{x}_i, z_i) &\sim \mathcal{N}\left(\alpha_{\text{hc}} + \beta_{\text{hc}}\mathbf{x}_i + \psi_{j[i]} + \tau_{\text{hc}} z_i + \omega_{j[i],\text{hc}} z_i, \sigma^2_{\text{hc},z}\right),
\end{aligned}
$$

where the variance terms for the two complier groups under treatment are assumed to be equal, $\sigma^2_{\text{cc}\,1} = \sigma^2_{\text{hc}\,1}$, and the random effects for site, $\{\psi_j\}$, are constrained to be equal across principal strata. The site-level estimates follow a multivariate Normal distribution:

$$
\begin{pmatrix} \psi_j \\ \omega_{j,\text{cc}} \\ \omega_{j,\text{hc}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \gamma \mathbf{w}_j \\ 0 \\ 0 \end{pmatrix}, \Sigma \right)
$$

where $\mathbf{w}_j$ is a vector of site-level covariates and Σ is an unconstrained covariance matrix.

B.5 Estimating the Principal Score Model

To illustrate principal score estimation, it is useful to start with the simpler instrumental variables case—that is, for binary D^*.

First, under one-sided noncompliance (i.e., only Compliers and Never Takers), principal score estimation proceeds almost identically to propensity score estimation in the usual observational setting. In particular, due to randomization, the principal score in the treatment group is the same as the overall principal score: $\mathbb{P}[S_i = s \mid Z_i = 1, \mathbf{X}_i = \mathbf{x}] = \mathbb{P}[S_i = s \mid \mathbf{X}_i = \mathbf{x}]$. Since we can directly observe compliance type in the treatment group, we simply estimate a model using covariates to predict compliance type among treated units. See Hill et al. (2002) and Jo and Stuart (2009).

Second, under two-sided noncompliance (i.e., Compliers, Always Takers, and Never Takers), we observe mixtures of compliance types under both treatment and control, which complicates estimation. One approach is to proceed non-parametrically, using kernel density estimation to estimate $(\pi_{a|\mathbf{x}}, \pi_{c|\mathbf{x}}, \pi_{n|\mathbf{x}})$. In high dimensions, however, this is impractical. Instead, we recommend a simple data augmentation strategy, due to Ibrahim (1990) and applied to causal inference by, among others, Aronow and Carnegie (2013) and Hsu and Small (2014). The essential idea is to use the same model as in Section B.3, except without any outcomes:

- **Estimate the principal score.** Given the vector of compliance types, estimate the principal score via multinomial logistic regression, ignoring treatment assignment.

- **Impute compliance type.** Given the principal score model, impute compliance types for all individuals with unknown type.

The principal score approach for the full model (i.e., multi-valued D) proceeds exactly as under the two-sided noncompliance setting. In addition, we extend the multinomial logistic regression to account for center-level variation, as in Appendix B.3.

B.6 Estimating Quantile Treatment Effects

We provide a brief overview of our approach for estimating the quantile treatment effects (QTE) by principal stratum.

We begin with the simpler setting without covariates. In this case, the overall QTE is simply the differences in observed quantiles under treatment and control. The QTE for Compliers, also known as IV-QTE, is

$$\tau_{\text{co}}(q) = G^{-1}_{\text{co}\,1}(q) - G^{-1}_{\text{co}\,0}(q),$$

where $G^{-1}_{sz}(q)$ is the qth quantile for a given stratum outcome distribution, $g_{sz}(y)$. As with the main results, there are two basic approaches for estimating $\tau_{\text{co}}(q)$. In the moment-based approach, we non-parametrically estimate $G^{-1}_{\text{co}1}(q)$ and $G^{-1}_{\text{co}0}(q)$ and subtract (Imbens and Rubin, 1997b; Abadie et al., 2002). In the Bayesian model-based approach with Normal components, we obtain posterior estimates for μ_{sz} and σ^2_{sz}. We then calculate the posterior predictive distribution for the QTE via the Normal quantile function, $\Phi^{-1}(q)$.

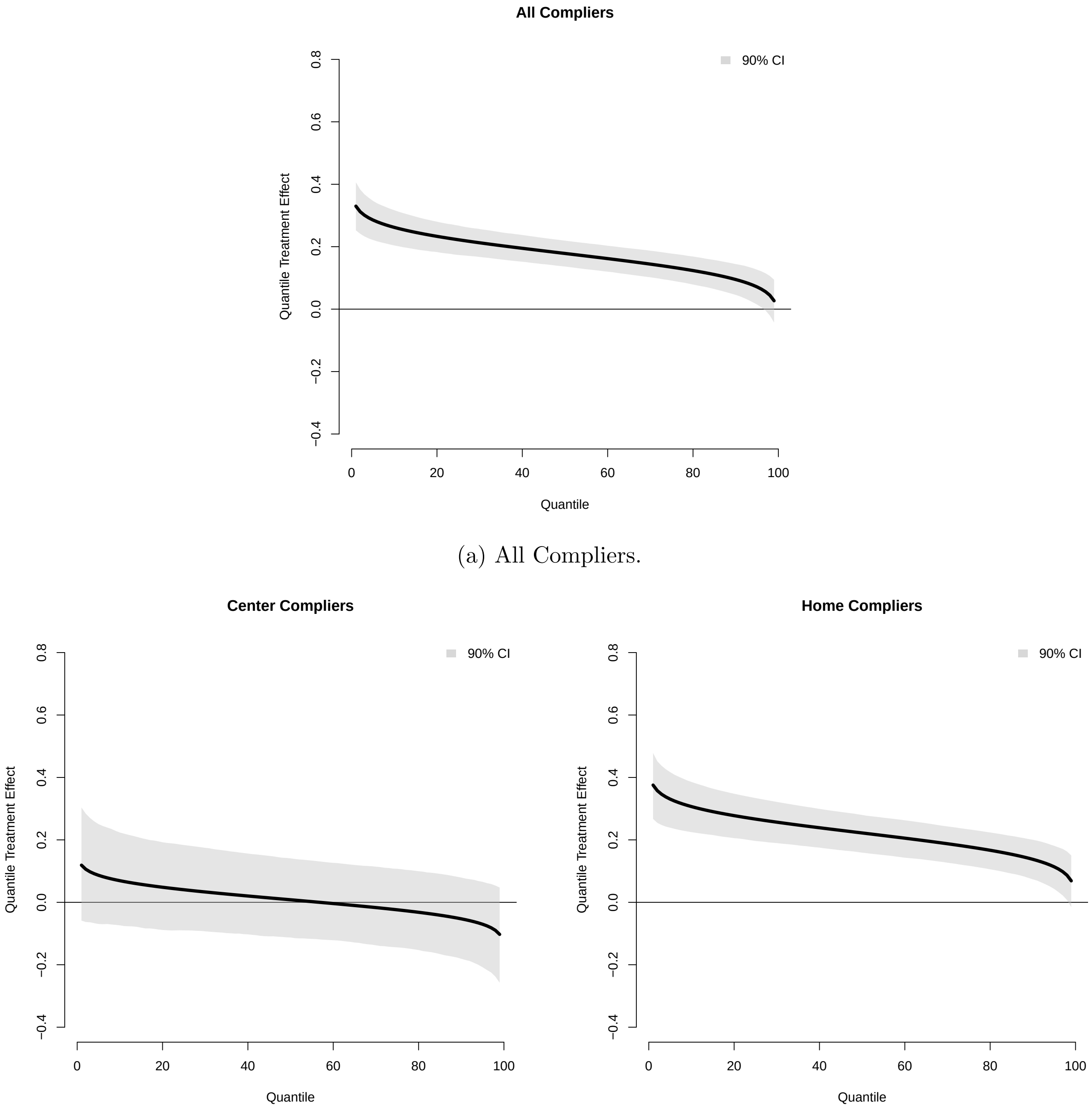
The QTE for principal strata, PS-QTE, are:

$$\tau_{\mathrm{cc}}(q) = G^{-1}_{\mathrm{cc}\,1}(q) - G^{-1}_{\mathrm{cc}\,0}(q)$$
$$\tau_{\mathrm{hc}}(q) = G^{-1}_{\mathrm{hc}\,1}(q) - G^{-1}_{\mathrm{hc}\,0}(q).$$

As with estimating main effects, a moment-based approach is no longer possible. However, we can still use the full principal stratification model to obtain posterior estimates for μ_{sz} and σ^2_{sz} and then use the Normal quantile function to calculate the posterior predictive distibution for the two QTEs. Alternatively, we can use the same stratification model to impute stratum membership for each MCMC iteration. We then directly estimate the PS-QTE for all Center Compliers and Home Compliers for a given MCMC iteration (i.e., by directly estimating the quantiles among all Center Compliers at that iteration), and combine the resulting estimates across iterations. While this second approach does not depend as directly on the assumption of Normality, the estimates are still sensitive to the parametric model, especially in the tails.

Estimation is more complicated with covariates. In particular, there are two possible objects of interest: conditional and unconditional QTEs. For the conditional QTE, the impacts are on the outcome distributions conditional on covariates. Estimation of the conditional QTE is straightforward via quantile regression (e.g., Abadie et al., 2002; Angrist and Pischke, 2008). For the unconditional QTE, the impacts are on the marginal outcome distributions. Firpo (2007) proposes to estimate these effects via the difference between weighted treatment and control quantiles, with inverse propensity score weights. Frölich and Melly (2013) proposes a slightly different set of weights to estimate the unconditional IV-QTE. Following Bitler et al. (2014), we focus on the unconditional QTE here.

As in the no covariates case, there are two main approaches. First, we can obtain posterior estimates for all the principal stratification model parameters, directly compute the posterior predictive distribution at each value of the covariates, and then combine. Figure A1 shows these results for the Head Start year for all Compliers, Center Compliers, and Home Compliers. Second, Second, we can use the model to impute stratum membership for each MCMC iteration. We can then use the inverse propensity score weighting approach of Firpo (2007) to estimate the covariate-adjusted unconditional IV-QTE and PS-QTE for each MCMC draw, combining estimates across iterations for final inference. As in the no covariate case, this approach relies less on the Normality assumption, but is still likely to be sensitive to the particular model choice. Figure 4 shows comparable estimates using the inverse propensity score weighting approach. As expected, the results are very similar between the two approaches, though the impacts at the bottom of the distribution are larger without relying directly on the Normal quantiles.



(a) All Compliers.

(b) Center Compliers.

(c) Home Compliers.

Figure A1: Quantile treatment estimates by compliance type for the Head Start year using the Normal approximation, with approximate 90 percent credible intervals.

B.7 Posterior Predictive Checks

Following Rubin et al. (1984) and Gelman et al. (2013), we use posterior predictive checks to assess the fit of our full model to the observed data. Formally, let y be the observed data and θ be the parameter vector. Define y^{rep} as the replicated data that could have been observed if the study were replicated with the same model and the same value of θ that produced y. We can estimate the distribution of y^{rep} via the posterior predictive distribution,

$$p(y^{\text{rep}} \mid y) = \int p(y^{\text{rep}} \mid \theta) p(\theta \mid y) d\theta.$$

The intuition is to assess whether the replicated data produced from the model are similar to the observed data.

First, we can visually compare the overall distributions of y and y^{rep}. Figure A2 shows a histogram of the observed data, y, and five replicated data sets, y^{rep}, for PPVT score. The histograms are indistinguishable from each other, suggesting that the model correctly captures the main features of the outcome distribution.

Second, we compare specific features of the distributions for y and y^{rep}. Define $T(y)$ as a test statistic that only depends on the data. We then determine whether the observed value of the test statistic, $T(y)$, is similar to the test statistics for the replicated data, $T(y^{\text{rep}})$. We can assess this numerically via a posterior predictive p-value:

$$p_{\text{pp}} = \mathbb{P}\{T(y^{\text{rep}}) \geq T(y) \mid \text{data}\}.$$

Intuitively, this is the proportion of replicated test statistics that are more extreme than the observed test statistic. We can also assess this discrepancy using visual summaries.

A key issue is choosing an appropriate test statistic. Following Barnard et al. (2003) and Mattei et al. (2013), we use the following three test statistics, which aim to assess whether the model captures broad features of the signal and noise available in the data:

- $T_{\text{signal}}(y) = |\overline{Y}_{s1} - \overline{Y}_{s0}|$, where $\overline{Y}_{sz}$ is the outcome mean for all children in stratum s and condition z
- $T_{\text{noise}}(y) = \sqrt{\frac{s^2_{s1}}{N_{s1}} + \frac{s^2_{s0}}{N_{s0}}}$, where s^2_{sz} is the outcome variance for all children in stratum s and condition z
- $T_{\text{ratio}}(y) = \frac{T_{\text{signal}}(y)}{T_{\text{noise}}(y)}$

We then compare $T(y^{\text{rep}})$ to $T(y)$ for each test statistic and each stratum to compute a posterior predictive p-value, where values close to 0 or 1 indicate poor model fit. As shown in Table A1, all posterior predictive p-values are away from the extremes, showing excellent model fit overall.

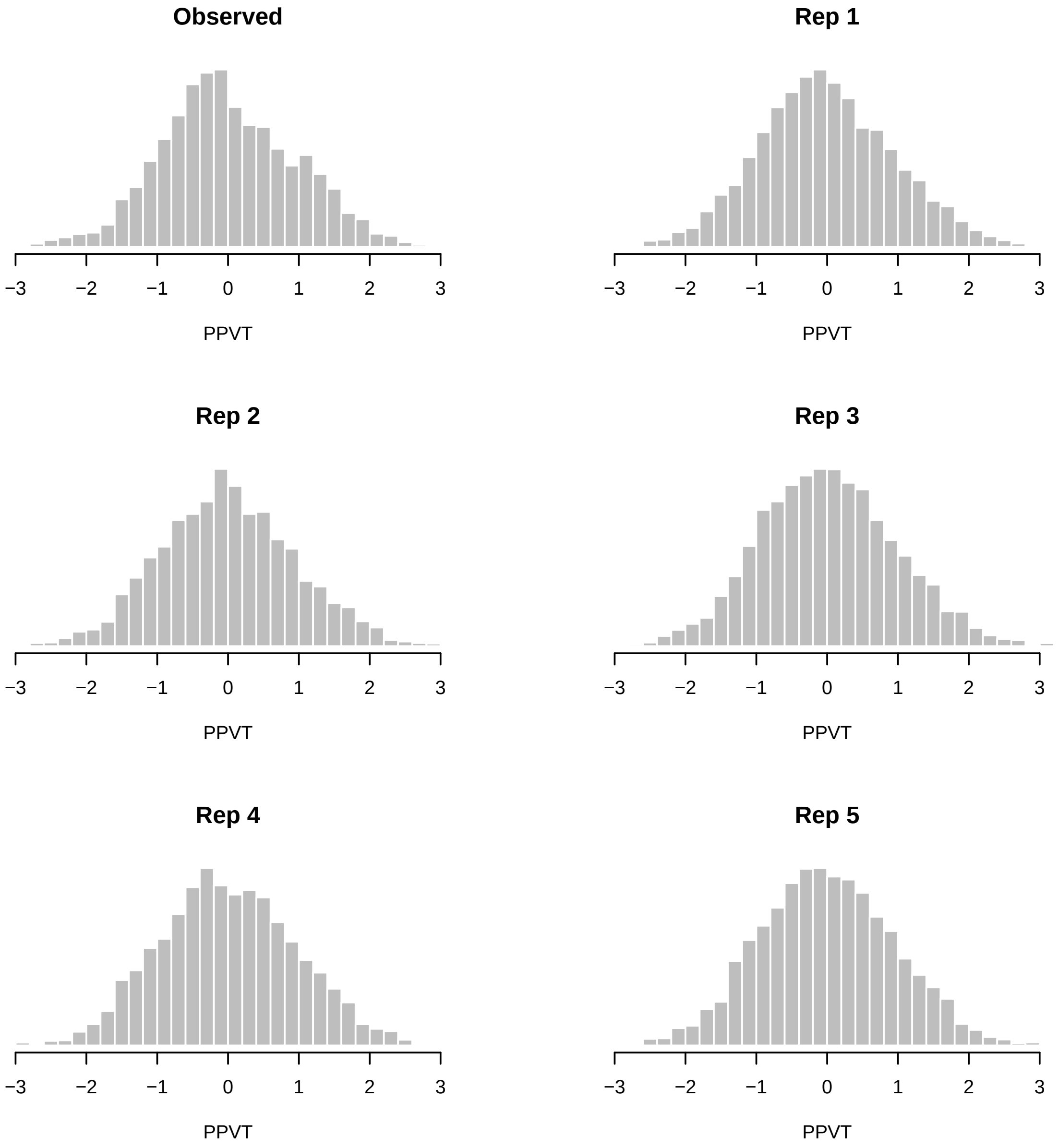


Figure A2: Overall outcome distributions for the observed data, y, and five replicated data sets, y^{rep}, using PPVT score.

Table A1: Posterior Predictive p-values

Stratum	$T_{\text{signal}}(y)$	$T_{\text{noise}}(y)$	$T_{\text{ratio}}(y)$
Always Head Start	0.77	0.41	0.77
Always Center-based	0.21	0.38	0.22
Always Home-based	0.40	0.20	0.41
Center-based Compliers	0.59	0.48	0.60
Home-based Compliers	0.40	0.09	0.46

C Missing Data

Survey nonresponse and general data missingness present major hurdles in the analysis of the Head Start Impact Study. In general, we account for three types of missing data in our analysis: missing outcomes, missing care type, and missing covariates. We focus in particular on missing pre-test scores, since this variable presents a unique set of challenges.

C.1 Missing Outcomes

As shown in Table A2, a substantial proportion of test scores scores are missing, with differential missingness by treatment status. Moreover, this missingness pattern is non-monotone; for example, some children have missing outcomes in the second year of observation but observed outcomes in the third year.

Table A2: Percent Missing PPVT Score.

	Control	Treatment	Difference
Pre-Test	0.33	0.19	-0.13
HS Year	0.24	0.13	-0.11
Pre-K/K	0.24	0.16	-0.08
K/1st	0.27	0.19	-0.08

Failing to account for such missingness, especially the differential missingness across experimental conditions, can lead to biased estimates (Frangakis and Rubin, 1999). A common approach to address this issue is to assume that outcomes are Missing at Random (MAR) (Rubin, 1976):

$$M_i \perp\!\!\!\perp Y_i \mid \mathbf{X}_i, Z_i, D_i^{obs} \qquad \text{(Missing at Random)}$$

where M_i is an indicator for missing outcome. We can re-write this as: $\mathbb{P}\{M_i \mid Y_i, \mathbf{X}_i, Z_i, D_i^{obs}\} = \mathbb{P}\{M_i \mid \mathbf{X}_i, Z_i, D_i^{obs}\}$. In other words, given covariates, treatment assignment, and observed child care setting, missing outcomes are just as likely to be low test scores as high test scores. This is a very sensible assumption in the Head Start Impact Study, as the data collection procedure depended heavily on the child's care setting.

Although implicit, this is the assumption behind the nonresponse adjustment in the official

HSIS report. Specifically, Puma et al. (2010a) estimate a model predicting item nonresponse as a function of child and family covariates, geography, Head Start offer, whether the child enrolled in Head Start, and other missingness indicators (e.g., missing pre-test score). The authors then weight units by the inverse of the predicted probability of nonresponse See the main text for additional discussion of weighting in HSIS. We make the same key assumption but implement this via the likelihood rather than via nonresponse weights (Mealli et al., 2004; Frumento et al., 2012). In that case, the missingness mechanism is *ignorable* (Rubin, 1976) since the likelihood factors such that the distribution of missingness indicators can be ignored in subsequent estimation. For computational reasons, we explicitly impute the missing outcomes rather than simply drop these terms from the likelihood, though the parameter estimation is the same.

Finally, we note that other missingness mechanisms are possible here. One promising but more technical assumption is *Latent Ignorability* (Frangakis and Rubin, 1999):

$$M_i \perp\!\!\!\perp Y_i \mid \mathbf{X}_i, Z_i, S_i. \qquad \text{(Latent Ignorability)}$$

Here, the missingness mechanism depends not only on observed care type, but also on the child's principal stratum. In other words, for control group children in center care, the probability of missingness could differ for Center Compliers and Always Center children. Nonetheless, there is no reason to believe that this relaxation is necessary here.

C.2 Missing Care Type

Based on survey responses alone, approximately 15 percent of children are missing information on their focal care setting. Using information elsewhere in the data, we can re-classify around one-third of these children. First, we utilize administrative data for "no shows" and "crossovers" collected for purposes of determining the IV estimate, which re-classifies 118 children. Second, we utilize Fall 2002 survey responses for parents who did not respond to the Spring 2003 survey, which re-classifies an additional 99 children. Table A3 shows detailed information for this procedure.

Table A3: Missing Care Type

	Raw Survey	+ Admin. Data	+ Fall 2002 Survey
Head Start	2,083	2,201	2,202
Other Center-Based Care	634	634	654
Other Home-Based Care	1,014	1,014	1,092
Unknown	654	536	437

Even after this effort, there is still substantial substantial missingness: with 16 percent of control group children and 6 percent of treatment group are missing care type, as shown in Table 2. Following Frumento et al. (2012), we assume that children with missing care type would belong to one of the otherwise existing principal strata, which means that children with missing care type are simply added to the likelihood components for these other strata. These relationships are shown in

Table A4. Importantly, we assume that children with missing care type could be in center care or home care, but not in Head Start; this is sensible since HSIS staff kept exhaustive records of Head Start attendance. Therefore, treatment group children with missing care type could be in either the Always Center or Always Home strata, but could not be Compliers, with the probability of being in the Always Center vs. Always Home stratum dependent on covariates and the outcome.

Table A4: Expanded version of Table 5b: Relationship between care type and principal stratum

Z_i	D_i^{obs}	Possible Principal Strata
1	HS	Always Head Start, Center Complier, Home Complier
1	Center	Always Center
1	Home	Always Home
1	?	Always Center, Always Home
0	HS	Always Head Start
0	Center	Always Center, Center Complier
0	Home	Always Home, Home Complier
0	?	Always Home, Home Complier, Always Center, Center Complier

C.3 Missing Covariates

As with outcomes and care setting, many children are missing a number of key covariates. The public use file for HSIS imputes these missing covariates, primarily via hot deck imputation. Importantly, the observed “donor cases” chosen for the hot deck were selected not only on the basis of covariates available in the HSIS file, but also on geographic and other programmatic variables not available to the public.

Ideally, we would multiply impute missing covariates alongside missing outcomes and missing care type, such as in Frumento et al. (2012). However, given the computational demands of multiple imputation and the non-public factors utilized for hot deck imputation, we use the imputed variables in the public use files. Note that this ignores the uncertainty associated with this imputation, though this uncertainty is likely quite small.

Finally, two of the 340 Head Start centers were missing all geocoded data. Note that while these were coded in the data file as centers, these are actually two of the grantees for which randomization was performed at the grantee level, rather than at the center/center group level. Given the small proportion of missingness here, we use simple mean value imputation to create a complete center-level data file.

C.4 Missing Pre-Test

The pre-test introduces several complications into the HSIS analysis. First, the pre-test is not, in fact, a true pre-test, since the test was conducted up to halfway through the Head Start year. However, we can reasonably assume the tests were administered early enough in the year that there was no meaningful treatment effect. The observed pre-tests are consistent this assumption.

Second, missingness is substantially higher in the control group than in the treatment group. As we do with missing outcomes, we multiply impute missing pre-test scores under the assumption of Missingness at Random:

$$M_i \perp\!\!\!\perp Pre_i \mid \mathbf{X}_i, Z_i, D_i^{obs}.$$

While this is a reasonable assumption for the missingness mechanism, the resulting imputations are conditional on D^{obs}. As a result, we cannot simply include these imputed scores in a later outcome regression, since the dependence on a post-treatment outcome would induce bias. At the same time, both M and Pre show large differences by observed care type, which suggests that dropping the dependence on D^{obs} altogether is not appropriate. Fortunately, we can sidestep this issue by conditioning on principal stratum membership, $S_i \equiv (D_i(0), D_i(1))$, as well as $\mathbf{X}$. Formally, we can factor the joint distribution of missingness and pre-test as follows under the MAR assumption:

$$\begin{aligned}\mathbb{P}\{M_i, Pre_i \mid \mathbf{X}_i, Z_i, D_i^{obs}\} &= \mathbb{P}\{M_i \mid \mathbf{X}_i, Z_i, D_i^{obs}\} \cdot \mathbb{P}\{Pre_i \mid \mathbf{X}_i, Z_i, D_i^{obs}\} \\ &= \mathbb{P}\{M_i \mid \mathbf{X}_i, Z_i, D_i^{obs}\} \cdot \mathbb{P}\{Pre_i \mid \mathbf{X}_i, S_i\}\end{aligned}$$

Ideally, we would incorporate the uncertainty in the imputation by drawing a new value of Pre_i for every MCMC iteration, using an imputation model that does not condition on either Z or Y. This is difficult in practice, however, especially with Stan. As a result, we adopt a hybrid approach similar to Frumento et al. (2012). First, we generate five separate data sets, which are identical except for different draws of the imputed pre-test score. Second, we run separate MCMC chains on each data set and combine these for final inference. Overall, we find very good convergence with this approach.

D Proofs

Proof of Lemma 1 (Non-Parametric Identification of the Distribution of Principal Strata).

The proof below is a simple extension of Lemma 3.1 in Abadie (2003). Under monotonicity and valid randomization, we have the following series of equalities, relating oberved population proportions to principal strata proportions. We repeatedly condition on X to emphasize the role of covariates, though for complete randomization, this equality holds unconditionally.

$$\begin{aligned}
\mathbb{P}\{D_1 = \text{HS} \mid Z = 1, X\} &= \mathbb{P}\{S = \text{Always Head Start} \mid X\} \\
&\quad + \mathbb{P}\{S = \text{Center Complier} \mid X\} \\
&\quad + \mathbb{P}\{S = \text{Home Complier} \mid X\} \\
\mathbb{P}\{D_1 = \text{Center} \mid Z = 1, X\} &= \mathbb{P}\{S = \text{Always Center} \mid X\} \\
\mathbb{P}\{D_1 = \text{Home} \mid Z = 1, X\} &= \mathbb{P}\{S = \text{Always Home} \mid X\}
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}\{D_0 = \text{HS} \mid Z = 0, X\} &= \mathbb{P}\{S = \text{Always HS} \mid X\} \\
\mathbb{P}\{D_0 = \text{Center} \mid Z = 0, X\} &= \mathbb{P}\{S = \text{Always Center} \mid X\} \\
&\quad + \mathbb{P}\{S = \text{Center Complier} \mid X\} \\
\mathbb{P}\{D_0 = \text{Home} \mid Z = 0, X\} &= \mathbb{P}\{S = \text{Always Home} \mid X\} \\
&\quad + \mathbb{P}\{S = \text{Home Complier} \mid X\}
\end{aligned}$$

We can then immediately identify the following:

$$\begin{aligned}
\mathbb{P}\{S = \text{Always HS} \mid X\} &= \mathbb{P}\{D^{obs} = \text{HS} \mid Z = 0, X\} \\
\mathbb{P}\{S = \text{Always Center} \mid X\} &= \mathbb{P}\{D^{obs} = \text{Center} \mid Z = 1, X\} \\
\mathbb{P}\{S = \text{Always Home} \mid X\} &= \mathbb{P}\{D^{obs} = \text{Home} \mid Z = 1, X\}
\end{aligned}$$

Substituting these into the first set of equations:

$$\begin{aligned}
\mathbb{P}\{S = \text{Center Complier} \mid X\} &= \mathbb{P}\{D^{obs} = \text{Center} \mid Z = 0, X\} - \mathbb{P}\{S = \text{Always Center} \mid X\} \\
&= \mathbb{P}\{D^{obs} = \text{Center} \mid Z = 0, X\} - \mathbb{P}\{D^{obs} = \text{Center} \mid Z = 1, X\} \\
\mathbb{P}\{S = \text{Home Complier} \mid X\} &= \mathbb{P}\{D^{obs} = \text{Home} \mid Z = 0, X\} - \mathbb{P}\{S = \text{Always Home} \mid X\} \\
&= \mathbb{P}\{D^{obs} = \text{Home} \mid Z = 0, X\} - \mathbb{P}\{D^{obs} = \text{Home} \mid Z = 1, X\}
\end{aligned}$$

Therefore, the principal score, $\mathbb{P}\{S = s \mid X\}$, is non-parametrically identified for all principal strata, s.

Proof of Lemma 2 (Distribution of Covariates by Principal Stratum). The proof is immediate. We want to identify the quantity $\mathbb{P}\{\mathbb{X}_i = \mathbf{x} \mid S_i = s\}$. From Bayes' Rule:

$$\mathbb{P}\{\mathbb{X}_i = \mathbf{x} \mid S_i = s\} = \frac{\mathbb{P}\{S_i = s \mid \mathbb{X}_i = \mathbf{x}\} \cdot \mathbb{P}\{\mathbb{X}_i = \mathbf{x}\}}{\mathbb{P}\{S_i = s\}}$$

From Lemma 1, we can non-parametrically identify $\mathbb{P}\{S_i = s\}$. Due to randomization, we can also identify $\mathbb{P}\{S_i = s \mid \mathbb{X}_i = \mathbf{x}\}$. Finally, we can identify the overall distribution of covariates in the sample, $\mathbb{P}\{\mathbb{X}_i = \mathbf{x}\}$. Therefore, $\mathbb{P}\{\mathbb{X}_i = \mathbf{x} \mid S_i = s\}$ is non-parametrically identified.

See Abadie (2003) for a similar argument.