*Article*

*Journal of Educational and Behavioral Statistics*
*2017, Vol. 42, No. 6, pp. 726–758*
*DOI: 10.3102/1076998617719726*


# Principal Score Methods: Assumptions, Extensions, and Practical Considerations

**Avi Feller**
*UC Berkeley*

**Fabrizia Mealli**
*Università di Firenze*

**Luke Miratrix**
*Harvard Graduate School of Education*

*Researchers addressing posttreatment complications in randomized trials often turn to principal stratification to define relevant assumptions and quantities of interest. One approach for the subsequent estimation of causal effects in this framework is to use methods based on the "principal score," the conditional probability of belonging to a certain principal stratum given covariates. These methods typically assume that stratum membership is as good as randomly assigned, given these covariates. We clarify the key assumption in this context, known as principal ignorability, and argue that versions of this assumption are quite strong in practice. We describe these concepts in terms of both one- and two-sided noncompliance and propose a novel approach for researchers to "mix and match" principal ignorability assumptions with alternative assumptions, such as the exclusion restriction. Finally, we apply these ideas to randomized evaluations of a job training program and an early childhood education program. Overall, applied researchers should acknowledge that principal score methods, while useful tools, rely on assumptions that are typically hard to justify in practice.*

Keywords: *principal stratification; principal score; noncompliance; causal inference*

## 1. Introduction

Principal stratification has become an increasingly important framework for addressing posttreatment complications in randomized trials in education, behavioral science, and related fields. Examples of principal stratification include estimating impacts in the presence of complex patterns of noncompliance, estimating how the impact of alternative high schools varies by the quality of the counterfactual school, and estimating the effect of a job training program on wages for individuals who would be employed regardless of the program (see Page, Feller, Grindal, Miratrix, & Somers, 2015, for a recent review).

Although principal stratification has gained widespread use for defining estimands of interest, the method of estimation can differ dramatically from application to application. In the article first defining principal stratification, for example, Frangakis and Rubin (2002) advocate a full model-based estimation strategy such as that found in Imbens and Rubin (1997) and Hirano, Imbens, Rubin, and Zhou (2000). Although this strategy is relatively common in statistics and biostatistics, there has been limited adoption of this approach among education and policy researchers, perhaps due to the complexity of implementation and unfamiliarity with Bayesian and likelihood methods. In this article, we explore an alternative approach that leverages covariates and various conditional independence assumptions to identify target estimands of interest. In particular, this approach utilizes the *principal score* (Hill, Waldfogel, & Brooks-Gunn, 2002), defined as the conditional probability of belonging to a certain principal stratum given covariates, which plays a role analogous to the (generalized) propensity score in traditional observational studies.

Our article makes two main contributions. Our first main contribution is to review existing principal score methods and clarify their corresponding assumptions. In particular, we investigate the role of the critical *principal ignorability* (PI) assumption and discuss two versions of this assumption, which we term strong and weak PI. We describe these assumptions in terms of real-world applications and argue that (1) strong PI is highly unlikely to hold in practice and that (2) weak PI, while strictly weaker, is still quite strong. We then discuss estimation via principal score weighting (Ding & Lu, 2017; Stuart & Jo, 2015), and introduce the key ideas with a simple illustration using a single, binary covariate.

Our second main contribution is to show how researchers can "mix and match" weak PI with other assumptions, especially the more common exclusion restriction. Thus, PI can be part of a broader menu of options for researchers seeking to estimate principal causal effects. This novel methodological contribution sets up the use of PI in more complex principal stratification settings.

We illustrate the key concepts with two randomized evaluations of social policy interventions with noncompliance. First, we review the example in Jo and Stuart (2009) of the Job Search Intervention Study (JOBS II), a randomized evaluation of a job training program for unemployed individuals that focused on mental health and job search skills (see also Mattei, Li, & Mealli, 2013). We use this example to highlight principal score methods in the simpler setting of one-sided noncompliance, in which only individuals assigned to the program are able to participate (i.e., "no crossovers"). As in Jo and Stuart (2009), we also fail to find evidence against the exclusion restriction for Never Takers in this evaluation.

Second, we explore the Head Start Impact Study (HSIS), a randomized evaluation of the Head Start program (Puma, Bell, Cook, Heid, & Shapiro, 2010).

This is an example of two-sided noncompliance, where control group children who were formally denied access to the Head Start program were nonetheless able to enroll. Although we are most interested in the impact of enrolling in Head Start on children's test scores, there is a concern that the assumptions necessary for a standard instrumental variables approach, namely the exclusion restriction, might be invalid (Gibbs, Ludwig, & Miller, 2011). Leveraging the mix and match strategy, we find evidence that there are positive impacts for Always Takers, although the uncertainty intervals are wide.

Overall, we believe that our article is a useful contribution to the small but growing literature on principal score methods. Like many statistical concepts, the idea of the principal score has multiple origins in different subfields. In biostatistics, the concept was first formalized by Follmann (2000), who called this the *compliance score* (see also Aronow & Carnegie, 2013; Joffe, Ten Have, & Brensinger, 2003). In the literature on statistics in the social sciences, the idea is due to Hill, Waldfogel, and Brooks-Gunn (2002), who introduced the term *principal score* (see also Jo, 2002; Jo & Stuart, 2009; Stuart & Jo, 2015). There have been many examples of this approach in practice, particularly in education and program evaluation, with some recent prominent examples from Schochet and Burghardt (2007) and Zhai, Brooks-Gunn, and Waldfogel (2014) as well as variations in economics, especially Crépon, Devoto, Duflo, and Parienté (2015). Schochet, Puma, and Deke (2014) offer a recent overview. Porcher, Leyrat, Baron, Giraudeau, and Boutron (2016) give a recent simulation study. Ding and Lu (2017) provide theoretical justification for a more general setup and offer additional guidance on estimation and sensitivity analysis.

This article proceeds as follows. Section 2 defines the relevant estimands and assumptions in the case of one-sided noncompliance. Section 3 extends these ideas to the case of two-sided noncompliance. Section 4 defines principal scores and discusses some of their properties. Section 5 gives details for estimating causal effects using principal scores. Section 6 applies the underlying methods to JOBS II and HSIS. Section 7 offers some thoughts for future research and concludes. The Appendix includes a discussion of some alternative estimation methods, a comparison with other identifying assumptions, and a proof of some desirable properties of the principal score.

## 2. One-Sided Noncompliance: JOBS II

### *2.1. Setup and Estimands*

Following Jo and Stuart (2009), we illustrate the key concepts for principal score methods for one-sided noncompliance using the JOBS II, a randomized evaluation of a job training program for unemployed individuals that focused on mental health and job search skills. Let $Z_i \in \{0,1\}$ be an indicator for whether individual $i$ is randomly offered the opportunity to enroll in the program. Also, let

TABLE 1.
*Relationship Between Observed Groups and Principal Strata in the Job Search Intervention Study*

| $Z_i$ | $D_i^{\text{obs}}$ | Principal Strata |
|---|---|---|
| 1 | 1 | Complier |
| 1 | 0 | Never Taker |
| 0 | 0 | Complier or Never Taker |

$Y_i^{\text{obs}}$ denote individual $i$'s observed outcome of interest, which we will set as a measure of depression 6 months after randomization. We invoke the stable unit treatment value assumption, which states that there is no interference between units and that there is only one version of the treatment (see Imbens & Rubin, 2015). With this assumption, we can define the potential outcomes for individual $i$, $Y_i(1)$ and $Y_i(0)$, which denote the depression score if individual $i$ is assigned to treatment or control, respectively. Given this, a natural estimand is the overall intent to treat (ITT),

$$\text{ITT} = \mathbb{E}[Y_i(1) - Y_i(0)],$$

which is the average impact of the opportunity to enroll on depression.

An important complication is that only 59% of the individuals assigned to treatment actually participated in the program. Let $D_i \in \{0, 1\}$ be an indicator for whether individual $i$ participated in the program, defined as attending at least one session, with corresponding potential outcomes $D_i(1)$ and $D_i(0)$ indicating whether individual $i$ will participate if assigned to treatment or control, respectively. Since these are potential outcomes, we regard them as fixed pretreatment. In JOBS II, individuals assigned to control were unable to attend these sessions; thus, $D_i(0) = 0$ for all $i$. Following Angrist, Imbens, and Rubin (1996) and Frangakis and Rubin (2002), we define *compliance types* or *principal strata* based on the joint values of treatment received under treatment and control, $(D_i(0), D_i(1))$. Since $D_i(0) = 0$ for all individuals, principal strata are completely defined by $D_i(1)$. We then have two compliance types:

$$S_i \equiv \begin{cases} \text{Never Taker } (n) & \text{if } D_i(0) = 0 \text{ and } D_i(1) = 0, \\ \text{Complier } (c) & \text{if } D_i(0) = 0 \text{ and } D_i(1) = 1. \end{cases}$$

In addition, define $\pi_s \equiv \mathbb{P}\{S_i = s\}$ as the proportion of stratum $s$ in the population. We do not fully observe these strata. Instead, we observe groups based on treatment assignment and participation. Table 1 shows the relationship between observed groups, defined by $Z$ and $D^{\text{obs}}$, and (partially) latent compliance types.

Since the joint values $(D_i(0), D_i(1))$ are fixed for each individual, we can regard $S_i$ as a pretreatment covariate. Therefore, we can think of subgroup treatment effects for Never Takers and Compliers the same way we would consider subgroup effects among men and women or among old and young. Within each principal stratum, it is as if we have a randomized experiment that could allow us to estimate these *principal causal effects* of interest (Frangakis & Rubin, 2002):[1]

$$\text{CACE} = \mathbb{E}\{Y_i(1) - Y_i(0)|S_i = \text{c}\} = \mu_{c1} - \mu_{c0},$$

$$\text{NACE} = \mathbb{E}\{Y_i(1) - Y_i(0)|S_i = \text{n}\} = \mu_{n1} - \mu_{n0},$$

where $\mu_{sz} \equiv \mathbb{E}\{Y_i(z)|S_i = s\}$, CACE is the Complier average causal effect, and NACE is the Never Taker average causal effect. We are primarily interested in the CACE, which is the effect of actually participating in the job training program.

We have the following relationships between average outcomes for observed groups, $\overline{Y}_{zd} \equiv \mathbb{E}\{Y_i^{\text{obs}}|Z_i = z, D_i^{\text{obs}} = d\}$, and average outcomes for (partially) latent principal strata, $\mu_{sz}$:

$$\begin{aligned} \overline{Y}_{11} &= \mu_{c1}, \\ \overline{Y}_{10} &= \mu_{n1}, \\ \overline{Y}_{00} &= \pi_c\mu_{c0} + (1 - \pi_{\text{c}})\mu_{n0}. \end{aligned} \tag{1}$$

The above shows that, since we directly observe whether individuals assigned to treatment are Compliers or Never Takers, we can immediately estimate $\mu_{c1}$ and $\mu_{n1}$ via the average observed outcomes, $\overline{Y}_{11}$ and $\overline{Y}_{10}$, respectively. We can also directly estimate the proportion of Compliers $\pi_c$ by the proportion of individuals assigned to treatment who participate in the program. Due to randomization, the distribution of compliance types is the same in the treatment and control groups in expectation. Finally, $\pi_n = 1 - \pi_c$.

Unfortunately, estimating the corresponding principal stratum means for individuals assigned to control is more difficult. As shown in Equation 1, individuals assigned to control are a mixture of Compliers and Never Takers. Without additional structure, we have one equation and two unknowns and cannot identify $\mu_{c0}$ and $\mu_{n0}$.

Following Angrist et al. (1996), the classic solution to this problem is to assume the exclusion restriction for the Never Takers; that is, to assume that $\text{NACE} = \mu_{n1} - \mu_{n0} = 0$. This assumption is reasonable when we believe that the only impact of randomization on the outcome is via participating in the program. Thus, if the encouragement has no impact on program participation, the encouragement should also have no impact on the outcome. At the same time, this assumption rules out ("excludes") other possible effects of randomization on the outcome. In the JOBS II example, this assumption states that there is no impact of randomization on depression for those individuals who would never participate in the program; that is, there is no "placebo" or motivation effect.

Under the exclusion restriction, we can use the standard instrumental variable approach to directly estimate $\mu_{n1}$ from the treatment arm, which reduces the above equation to a single unknown. If the exclusion restriction does not hold, however, the resulting estimate for the CACE could be biased. The goal of a principal score analysis for JOBS II is to estimate the CACE without relying on the exclusion restriction.

Before explaining principal score analysis in detail, we briefly note that there are a range of alternative approaches that broadly fall under the umbrella of principal stratification. First, without any additional assumptions, the means in Equation 1 are set identified; that is, we can obtain nonparametric bounds for $\mu_{c0}$ and $\mu_{n0}$ rather than point estimates (e.g., Zhang & Rubin, 2003). Although unadjusted bounds are typically too wide for practical use, there are several strategies for sharpening these bounds, such as by leveraging pretreatment covariates (Grilli & Mealli, 2008; D. S. Lee, 2009; Long & Hudgens, 2013; Miratrix, Furey, Feller, Grindal, & Page, 2017) or secondary outcomes (Mealli & Pacini, 2013). Second, we could exploit specific conditional independence assumptions between covariates and outcomes conditional on principal strata (Ding, Geng, Yan, & Zhou, 2011; Mealli, Pacini, & Stanghellini, 2016) or among multiple outcomes conditional on principal strata (Mealli et al., 2016) to achieve full identification of principal causal effects. Jo (2002) uses such methods to estimate the CACE in the JOBS II example. The necessary assumptions are often quite strong in practice, however.

Finally, we could use a fully model-based estimation strategy, such as originally proposed by Imbens and Rubin (1997), which requires imposing distributional assumptions on the outcome to disentangle the mixture. For example, we might assume that $Y_i^{\text{obs}}|S_i = s, Z_i = z \sim N(\mu_{sz}, \sigma^2)$, possibly conditional on covariates. Mattei, Li, and Mealli (2013) use this approach to estimate the CACE in the JOBS II example. A key concern of these model-based approaches is that results are sensitive to the particular choice of model. Feller, Greif, Miratrix, and Pillai (2016) also highlight additional dangers of relying on parametric mixture models for estimating principal causal effects.

### *2.2. Principal Ignorability*

As in Jo and Stuart (2009), we observe a rich set of pretreatment covariates for each individual $i$, denoted by $\mathbf{x}_i$. In the JOBS II example, we focus on seven covariates: baseline measures of depression, sense of mastery, economic hardship, and motivation, as well as age, years of schooling, and gender. Intuitively, PI states that, given these covariates, whether an individual is a Complier or Never Taker is as good as randomly assigned. We now make this statement more precise, showing that there is a *strong* and *weak* version of PI. Consider two conditional independence assumptions:[2]

$$\mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}, D_i(1) = 1] = \mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}, D_i(1) = 0] = \mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}], \tag{2}$$

$$\mathbb{E}[Y_i(0)|\mathbf{X}_i = \mathbf{x}, D_i(1) = 1] = \mathbb{E}[Y_i(0)|\mathbf{X}_i = \mathbf{x}, D_i(1) = 0] = \mathbb{E}[Y_i(0)|\mathbf{X}_i = \mathbf{x}]. \tag{3}$$

Weak PI assumes Equation 3. Strong PI assumes both Equations 2 and 3. Note that $D_i(1) = 1$ and $D_i(1) = 0$ are equivalent to $S_i = c$ and $S_i = n$, respectively. We can also rewrite strong PI in terms of impacts:

$$\text{CACE}(\mathbf{x}) = \text{NACE}(\mathbf{x}) = \text{ITT}(\mathbf{x}), \tag{4}$$

where $\text{CACE}(\mathbf{x}) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}, D_i(1) = 1]$, $\text{NACE}(\mathbf{x}) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}, D_i(1) = 0]$, and $\text{ITT}(\mathbf{x}) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}]$, which are the subgroup CACE, NACE, and ITT estimands, respectively. In other words, given covariates, the CACE, NACE, and ITT are all equal.

Equation 2 states that, for individuals assigned to treatment, whether they *actually* participate in the program is unrelated to their outcome, given covariates. For illustration, imagine we observe a group of 30-year-old men with high school degrees and identical baseline measures of depression, sense of mastery, economic hardship, and motivation. All of these men are assigned to the JOBS II treatment group, but half attend the program and the other half do not. Equation 2 states that the average outcomes for these two groups should be the same, even though we know that half actually attended the training program and the other half did not. Equivalently, Equation 4 states that the subgroup $\text{CACE}(\mathbf{x})$ and $\text{NACE}(\mathbf{x})$ are equal to the overall $\text{ITT}(\mathbf{x})$, which precludes any impact from program participation. There are some specific cases in which this might be plausible, for example, if Never Takers receive an equivalent program to Compliers. But, in general, this assumption is quite difficult to justify in practice.

Equation 3 states that, for individuals assigned to control, whether they *would have participated in the program if offered* is unrelated to their outcome given covariates. This is a weaker condition than Equation 2 since $Y_i(0)$ and $D_i(1)$ are realized in different states of the world. Returning to JOBS II, imagine that the group of 30-year-old men is instead assigned to the control condition; we no longer know which half would participate in the program if offered (i.e., whether they are Compliers or Never Takers). Equation 3 states that, given covariates, knowing whether these men would participate *when assigned to treatment* is unrelated to their outcomes *when assigned to control*. Importantly, neither Compliers nor Never Takers enroll in the program when assigned to control, so they have the same observed program participation (i.e., $D_i^{\text{obs}} = 0$). Thus, while Equation 3 is still a very strong assumption, it is generally weaker than Equation 2 and is more plausible in the JOBS II setting.

As shown in Table 1, in the case of one-sided noncompliance, we directly observe stratum membership for individuals assigned to treatment. Thus, Equation 2 is unnecessary for identification since we can estimate the relevant mean outcomes directly. In fact, as we discuss below, we can compare these estimates to what we would get if this assumption were true, giving an immediate testable implication.

## 3. Two-Sided Noncompliance: Head Start Impact Study

### *3.1. Setup and Estimands*

We now extend the setup and assumptions from one-sided noncompliance to the more complex case of two-sided noncompliance, that is, where individuals assigned to the control arm have access to the program. Our running example is the Head Start Impact Study (Puma et al., 2010), discussed in more detail in Section 6.2. As above, let $Z_i$ denote whether child $i$ is randomly offered the opportunity to enroll in Head Start; $Y_i^{\text{obs}}$ denote child $i$'s observed outcome of interest, which we will set as the Peabody Picture Vocabulary Test (PPVT) score; and $\mathbf{x}_i$ denote a vector of pretreatment covariates including pretest score. Let $D_i \in \{0, 1\}$ be an indicator for whether child $i$ enrolls in Head Start. Importantly, children are classified as enrolling in Head Start regardless of which specific Head Start center they attend. Thus, many children formally denied a spot at their initial Head Start center (i.e., assigned to control) nonetheless enrolled in a different Head Start center; these children are denoted as $D_i = 1$.

Following Angrist et al. (1996), there are four possible compliance types (without additional restrictions):

$$S_i \equiv \begin{cases} \text{Always Taker } (a) & \text{if } D_i(0) = 1 \text{ and } D_i(1) = 1 \\ \text{Complier } (c) & \text{if } D_i(0) = 0 \text{ and } D_i(1) = 1 \\ \text{Defier } (d) & \text{if } D_i(0) = 1 \text{ and } D_i(1) = 0 \\ \text{Never Taker } (n) & \text{if } D_i(0) = 0 \text{ and } D_i(1) = 0. \end{cases}$$

In HSIS, it is reasonable to invoke the *monotonicity* or "no defiers" assumption, which assumes that the offer of enrollment in Head Start did not induce any children to *not* enroll in the program and vice versa (for discussion, see Puma et al., 2010). This yields three possible principal strata: $S_i \in \{a, c, n\}$. The primary estimand of interest is the CACE, the effect of enrolling in Head Start for those children who would enroll if offered the opportunity to do so and would not enroll if not offered. We are also interested in the Always Taker average causal effect (AACE), the effect of the offer of enrollment on children who would enroll in a Head Start center regardless of treatment assignment. For instance, this could be the impact of attending some Head Start centers that are higher quality than others. If this effect is nonzero, then the exclusion restriction for Always Takers is invalid and the standard instrumental variable estimate will be misleading.

Table 2 shows the relationship between observed groups and principal strata in this example. This yields four equalities:

TABLE 2.
*Relationship Between Observed Groups and Principal Strata in the Head Start Impact Study Under Monotonicity*

| $Z_i$ | $D_i^{\text{obs}}$ | Principal Strata |
|---|---|---|
| 1 | 1 | Complier or Always Taker |
| 1 | 0 | Never Taker |
| 0 | 1 | Always Taker |
| 0 | 0 | Complier or Never Taker |

$$\overline{Y}_{11} = \frac{\pi_c}{\pi_c + \pi_a}\mu_{c1} + \frac{\pi_a}{\pi_c + \pi_a}\mu_{a1},$$
$$\overline{Y}_{10} = \mu_{n1},$$
$$\overline{Y}_{01} = \mu_{a0},$$
$$\overline{Y}_{00} = \frac{\pi_c}{\pi_c + \pi_n}\mu_{c0} + \frac{\pi_n}{\pi_c + \pi_n}\mu_{n0}.$$

We can immediately estimate $\mu_{a0}$ and $\mu_{n1}$ via $\overline{Y}_{01}$ and $\overline{Y}_{10}$, the observed average outcomes for $\{i : Z_i = 0, D_i^{\text{obs}} = 1\}$ and $\{i : Z_i = 1, D_i^{\text{obs}} = 0\}$, respectively. Analogous to the one-sided case, we can also estimate the overall proportion of each principal stratum: $\pi_a = \mathbb{P}\{D_i(0) = 1|Z_i = 0\}$, $\pi_n = \mathbb{P}\{D_i(1) = 0|Z_i = 1\}$, and $\pi_c = 1 - \pi_a - \pi_a$.

However, we now have two mixtures to disentangle: the mixture of Compliers and Always Takers assigned to treatment and the mixture of Compliers and Never Takers assigned to control.

### *3.2. Principal Ignorability*

We now clarify the PI assumptions in the case of two-sided noncompliance and highlight the ability to mix and match PI and exclusion restriction assumptions.

*3.2.1. Strong PI.* First, we can extend Equations 2 and 3 to allow for three compliance types:

$$\mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}, S_i = a] = \mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}, S_i = n] = \mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}, S_i = c], \quad (5)$$

$$\mathbb{E}[Y_i(0)|\mathbf{X}_i = \mathbf{x}, S_i = a] = \mathbb{E}[Y_i(0)|\mathbf{X}_i = \mathbf{x}, S_i = n] = \mathbb{E}[Y_i(0)|\mathbf{X}_i = \mathbf{x}, S_i = c]. \quad (6)$$

Strong PI assumes both of these equations and is essentially identical to the one-sided case. Equivalently, $\text{CACE}(\mathbf{x}) = \text{NACE}(\mathbf{x}) = \text{AACE}(\mathbf{x}) = \text{ITT}(\mathbf{x})$. In the context of HSIS, these assumptions state that, given treatment assignment and covariates, whether a child actually attends Head Start is unrelated to that child's observed test score. Equivalently, given covariates, whether a child actually attends Head Start is unrelated to the impact of randomization on test score. Again, strong PI is quite a strong assumption and seems implausible in practice. As in the one-sided case, this yields testable implications since we directly observe the average outcome for Always Takers assigned to control and the average outcome for Never Takers assigned to treatment.

*3.2.2. Weak PI for Compliers, Always Takers, and Never Takers.* Similar to the one-sided case, we can relax strong PI, though we now need a pair of assumptions:

$$\mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}, D_i(1) = 1, D_i(0) = d] = \mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}, D_i(1) = 1] \qquad \text{for } d = 0, 1, \tag{7}$$

$$\mathbb{E}[Y_i(0)|\mathbf{X}_i = \mathbf{x}, D_i(0) = 0, D_i(1) = d] = \mathbb{E}[Y_i(0)|\mathbf{X}_i = \mathbf{x}, D_i(0) = 0] \qquad \text{for } d = 0, 1. \tag{8}$$

Equations 7 and 8 comprise weak PI in this setting. Equation 7 states that, for children assigned to treatment who enroll in Head Start, whether they *would have enrolled in Head Start if denied a spot* is unrelated to their test scores, given covariates. Analogously, Equation 8 states that, for children assigned to control who do not enroll in Head Start, whether they *would have enrolled in Head Start if offered* is unrelated to their test scores, given covariates. This equation is the weak PI assumption under one-sided noncompliance, in which $D_i(1)$ and $Y_i(0)$ are realized in different states of the world. Equations 7 and 8 are strictly weaker assumptions than Equations 5 and 6 since they only apply to a subset of individuals in each treatment arm, for example, Equation 5 applies to all children assigned to treatment while Equation 7 only applies to those children assigned to treatment who actually enroll in Head Start (i.e., excluding Never Takers). In addition, these equalities are only for units within observationally indistinguishable groups.

*3.2.3. Weak PI for Compliers and Always Takers and exclusion restriction for Never Takers.* For two-sided noncompliance, we need identifying assumptions for each of the two mixtures. This suggests a strategy where we target these assumptions to each mixture; that is, we mix and match PI and exclusion restrictions. For example, in the Head Start scenario, we can (1) assume that, given covariates, treatment outcomes are the same for Always Takers and Compliers; and (2) assume that there is no effect of the offer of enrollment on those children who would never enroll in Head Start regardless of treatment assignment. This yields,

$$\mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}, D_i(1) = 1, D_i(0) = d] = \mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}, D_i(1) = 1] \qquad \text{for } d = 0, 1, \tag{9}$$

$$\mathbb{E}[Y_i(1)|S_i = n] = \mathbb{E}[Y_i(0)|S_i = n]. \tag{10}$$

We have replaced the second weak PI assumption with an exclusion restriction. The exclusion restriction for Never Takers is relatively uncontroversial for HSIS since these children do not change their behavior as a result of randomization and never enroll in a Head Start program (Gibbs et al., 2011; Puma et al., 2010).

## 4. Principal Scores

### *4.1. Definition*

As in standard observational studies, researchers will typically find PI more plausible if they have a rich set of covariates. Without additional structure, however, high-dimensional $\mathbf{X}$ can be unwieldy in practice, as we would observe few units for any given combination of covariate values. Borrowing from the propensity score literature, we can reduce the dimensionality of $\mathbf{X}$ by calculating the principal score (Hill et al., 2002) for stratum $s$:

$$e_s(\mathbf{x}) \equiv \mathbb{P}\{S_i = s | \mathbf{X}_i = \mathbf{x}\},$$

the conditional probability that individual $i$ belongs to stratum $s$ given covariates $\mathbf{X}_i = \mathbf{x}$. With some abuse of notation, we will sometimes write this as $e_s(\mathbf{x}_i)$ to emphasize that this is the principal score for individual $i$.

Lemma 1 shows that principal scores share two desirable properties with propensity scores. A proof of this lemma is in Appendix A.3; see also Jo and Stuart (2009) and Ding and Lu (2017).

> **Lemma 1 (Properties of the Principal Score):** The principal score, $e_s(\mathbf{x})$, is a balancing score in the sense that $S_i \perp\!\!\!\perp \mathbf{X}_i | e_s(\mathbf{x})$. Furthermore, if either strong or weak PI holds given $\mathbf{X}_i$ that same assumption also holds given $e_s(\mathbf{x})$.

As a result, we can reduce the dimensionality of $\mathbf{X}$ to a scalar in an analogous way to the propensity score in observational studies. Similar to propensity scores, the balancing score property holds only in theory and must be assessed in practice. Section 5.4 discusses this in more detail in addition to assessing the overall principal score model fit.

### *4.2. Estimation*

*4.2.1. One-sided noncompliance.* In this setting, we can estimate the principal score directly since $\mathbb{P}\{D_i^{\text{obs}} = 1 | Z_i = 1, \mathbf{X}_i = \mathbf{x}\} = \mathbb{P}\{S_i = c | Z_i = 1, \mathbf{X}_i = \mathbf{x}\} = \mathbb{P}\{S_i = c | \mathbf{X}_i = \mathbf{x}\} = e_c(\mathbf{x})$. Therefore, we can obtain a nonparametric estimate of $e_{\mathrm{c}}(\mathbf{x})$ asymptotically by estimating the proportion of $D_i^{\text{obs}} = 1$ for each $\mathbf{X}_i = \mathbf{x}$ in the treatment group (see, e.g., Abadie, 2003). Alternatively, following Schochet and Burghardt (2007) and Jo and Stuart (2009), we can estimate $e_c(\mathbf{x})$ via modeling, such as with a (logistic) regression of $D$ on $\mathbf{X}$ among individuals with $Z_i = 1$. Regardless, once we have a model, we can estimate $e_c(\mathbf{x})$ for the entire sample including the control group. See Abadie, Chingos, and West (2016) for additional considerations in using the same data for estimating both the principal score model and the causal effect.

*4.2.2. Two-sided noncompliance: marginal method.* Estimation is slightly more complicated in this setting because we never observe Compliers directly. We can, however, estimate two separate, *marginal* models and combine. This approach takes advantage of the useful fact that, under monotonicity, we can directly observe Never Takers assigned to treatment and Always Takers assigned to control. In particular, we can directly estimate $e_a(\mathbf{x}) \equiv \mathbb{P}\{S_i = a|\mathbf{X}_i = \mathbf{x}\}$ via the predicted probability from a (logistic) regression of $D$ on $\mathbf{X}$ in the control group. Similarly, we can estimate $e_n(\mathbf{x}) \equiv \mathbb{P}\{S_i = n|\mathbf{X}_i = \mathbf{x}\}$ via 1 minus the predicted probability from a (logistic) regression of $D$ on $\mathbf{X}$ in the treatment group. Then, by construction, $\widehat{e}_c(\mathbf{x}) = 1 - \widehat{e}_a(\mathbf{x}) - \widehat{e}_n(\mathbf{x})$. Of course, we could replace logistic regression with nonparametric regression or similar estimation approaches.

*4.2.3. Two-sided noncompliance: joint method.* An obvious concern is that separately estimating $\widehat{e}_a(\mathbf{x})$ and $\widehat{e}_n(\mathbf{x})$ could lead to estimates for $\widehat{e}_c(\mathbf{x})$ that are outside $[0, 1]$. We can impose this constraint by jointly estimating the principal score models. One estimation approach is via data augmentation, with stratum membership as a partially observed variable (for additional details, see Ding & Lu, 2017, as well as Ibrahim, 1990; Aronow & Carnegie 2013; Hsu & Small 2014; Zhang, Rubin, & Mealli, 2009). The key idea is to alternate between two steps. Starting with an initial vector of compliance types, repeat the following steps until convergence:

- *Estimate the principal score*. Given the vector of compliance types, estimate the principal score via multinomial logistic regression of $S$ on $\mathbf{X}$, ignoring treatment assignment.
- *Impute compliance type*. Given the principal score model, impute compliance types for all individuals with unknown type. For expectation maximization, this is via maximization. For Markov chain Monte Carlo, this is via imputation.

In practice, researchers can first compute the simpler "marginal" principal score estimates and only proceed to the "joint" model if estimates are outside $[0, 1]$.

## 5. Estimating Impacts Under Principal Ignorability

Researchers have proposed a range of methods for estimating principal causal effects given PI. We focus on the weighting method of Stuart and Jo (2015) and Ding and Lu (2017), which is intuitive and straightforward to implement. Other methods that we do not discuss here include regression (Bein, 2015; Joffe, Small, & Hsu, 2007) and matching (Hill et al., 2002; Jo & Stuart, 2009). For further discussion, see Porcher et al. (2016), who conduct extensive simulation studies and find that weighting has slightly better finite sample performance than matching in practice. As these alternative methods are inherently driven by PI, we anticipate that the intuition we give for the weighting method should carry over.

Appendix A.1 explores the "discrete subgroup" method of Schochet and Burghardt (2007). We show that, even under strong PI, this method only yields unbiased estimates in certain degenerate cases. Nonetheless, we view this approach as a useful exploratory method; it is particularly promising when researchers are actually interested in predicted subgroups rather than the principal strata themselves.

### *5.1. Estimation With Single, Binary Covariate*

To build intuition for more complex methods, we first show how to estimate impacts under PI in the simplest nontrivial case: one-sided noncompliance in which we assume that PI holds given a single, binary covariate. For illustration, we use an indicator in JOBS II that is coded "female" or "male." Let $X$ be this covariate, $X_i \in \{m, f\}$; let $p(x) \equiv \mathbb{P}\{X_i = x\}$ be the proportion of individuals in the population with $X_i = x$; let $p(x|s) = \mathbb{P}\{X_i = x|S_i = s\}$ be the proportion of individuals with $X_i = x$ among those in stratum $s$; and let $e_s(x) \equiv \mathbb{P}\{S_i = s|X_i = x\}$ be the proportion of individuals in stratum $s$ among those with $X_i = x$.

*5.1.1. Estimating stratum characteristics.* First, we can obtain $p(x|s)$ via Bayes's rule:

$$p(x|s) = \mathbb{P}\{X_i = x|S_i = s\} = \frac{\mathbb{P}\{S_i = s|X_i = x\}\ \mathbb{P}\{X_i = x\}}{\mathbb{P}\{S_i = s\}} = \frac{e_s(x)\, p(x)}{\pi_s}.$$

Under one-sided noncompliance, we can estimate $\pi_s$ via the observed proportion of individuals assigned to the treatment group who participate in the program; we can estimate $e_s(x)$ with the corresponding proportion for the subgroup with $X_i = x$. Finally, we can directly observe the overall proportion, $p(x)$. We then plug these sample analogs into the above equation to estimate $p(x|s)$.

Separately for each value of $X$, we can estimate $\overline{Y}_{zd}(x)$, the average outcome for individuals with $X_i = x$ assigned to condition $Z_i = z$ with observed participation $D_i^{\text{obs}} = d$. We can also ignore program participation and estimate $\overline{Y}_{z\cdot}(x) \equiv \mathbb{E}\{Y_i^{\text{obs}}|Z_i = z, X_i = x\}$, the average outcome for all individuals with $X_i = x$ assigned to treatment condition $Z_i = z$, averaging over $D$.

*5.1.2. Estimating average control outcomes.* To estimate $\mu_{c0}$ and $\mu_{n0}$, we assume Equation 3, weak PI. We can rewrite this assumption more compactly as

$$\mu_{c0}(f) = \mu_{n0}(f) = \overline{Y}_{0\cdot}(f) \qquad \text{and} \qquad \mu_{c0}(m) = \mu_{n0}(m) = \overline{Y}_{0\cdot}(m),$$

where $\mu_{sz}(x) \equiv \mathbb{E}\{Y_i(z)|S_i = s, X_i = x\}$. This states, for example, that average outcomes for females assigned to control are unrelated to whether they would take up the treatment if offered. Under this assumption, $\overline{Y}_{0\cdot}(f)$, the average outcome for those with $X_i = f$ in the control group, is equal to both $\mu_{c0}(f)$ and $\mu_{n0}(f)$.

Given this, estimating the overall means for Compliers and Never Takers assigned to control is immediate. For Compliers, the overall average, $\mu_{c0}$, is the weighted mean of the average outcome for males, $\mu_{c0}(m)$, and the average outcome for females, $\mu_{c0}(f)$, weighted by the group sizes:

$$\begin{aligned}\mu_{c0} &= p(f|\mathrm{c})\ \mu_{c0}(f) + p(m|\mathrm{c})\ \mu_{c0}(m) \\ &= \frac{e_c(f)p(f)}{\pi_c}\ \overline{Y}_{0\cdot}(f) + \frac{e_c(m)p(m)}{\pi_c}\ \overline{Y}_{0\cdot}(m)\cdot\end{aligned} \tag{11}$$

The first line demonstrates that this is a weighted average of two subgroup means. The second line rewrites this weighted average in terms of quantities that we can directly estimate. Specifically, we have the following plug-in estimators:

$$\widehat{\mu}_{c0} = \frac{\widehat{e}_c(f)\ \widehat{p}(f)}{\widehat{\pi}_c}\ \widehat{Y}_{0\cdot}(f) + \frac{\widehat{e}_c(m)\ \widehat{p}(m)}{\widehat{\pi}_c}\ \widehat{Y}_{0\cdot}(m),$$

$$\widehat{\mu}_{n0} = \frac{\widehat{e}_n(f)\ \widehat{p}(f)}{\widehat{\pi}_n}\ \widehat{Y}_{0\cdot}(f) + \frac{\widehat{e}_n(m)\ \widehat{p}(m)}{\widehat{\pi}_n}\ \widehat{Y}_{0\cdot}(m),$$

where $\widehat{Y}_{zd}(x)$ is the sample average outcome for individuals with $Z_i = z$, $D_i^{\text{obs}} = d$, and $X_i = x$.

*5.1.3. Estimating average treatment outcomes.* There are two options for estimating the average stratum outcomes under treatment. First, we can directly estimate $\mu_{c1}$ and $\mu_{n1}$ without any additional assumptions:

$$\widehat{\mu}_{c1}^{\text{Weak PI}} = \widehat{Y}_{11}, \qquad \widehat{\mu}_{n1}^{\text{Weak PI}} = \widehat{Y}_{10}$$

Alternatively, under strong PI, we can, just as above, estimate these quantities via a weighted average of the subgroup averages for men and women, $\widehat{Y}_{1\cdot}(m)$ and $\widehat{Y}_{1\cdot}(f)$, ignoring the actual program participation for each group:

$$\widehat{\mu}_{c1}^{\text{Strong PI}} = \frac{\widehat{e}_c(f)\ \widehat{p}(f)}{\widehat{\pi}_c}\ \widehat{Y}_{1\cdot}(f) + \frac{\widehat{e}_c(m)\ \widehat{p}(m)}{\widehat{\pi}_c}\ \widehat{Y}_{1\cdot}(m),$$

$$\widehat{\mu}_{n1}^{\text{Strong PI}} = \frac{\widehat{e}_n(f)\ \widehat{p}(f)}{\widehat{\pi}_n}\ \widehat{Y}_{1\cdot}(f) + \frac{\widehat{e}_n(m)\ \widehat{p}(m)}{\widehat{\pi}_n}\ \widehat{Y}_{1\cdot}(m).$$

Importantly, because $\hat{\mu}_{s1}^{\text{Weak PI}}$ and $\hat{\mu}_{s1}^{\text{Strong PI}}$ are two distinct estimators of the same quantity, strong PI yields a testable implication. In particular, if the estimates obtained via the weak and strong PI assumptions are not equal up to sampling error, Equation 2 must not hold. This test does not inform us, however, as to whether the control (weak) side of the assumption is valid.

*5.1.4. Estimating impacts.* Finally, we estimate the CACE and NACE as the difference in estimated means:

$$\widehat{\text{CACE}}^{\text{PI}} = \widehat{\mu}_{c1}^{\text{PI}} - \widehat{\mu}_{c0} \qquad \widehat{\text{NACE}}^{\text{PI}} = \widehat{\mu}_{n1}^{\text{PI}} - \widehat{\mu}_{n0},$$

where the superscript PI denotes either weak or strong PI.

Under strong PI, we can also rewrite the CACE estimator as a weighted average of the ITT estimates for males and females:

$$\begin{aligned}\widehat{\text{CACE}}^{\text{Strong PI}} &= \left[\frac{\widehat{e}_{\text{c}}(f)\,\widehat{p}_f}{\widehat{\pi}_{\text{c}}}\ \widehat{Y}_{1\cdot}(f) + \frac{\widehat{e}_{\text{c}}(m)\,\widehat{p}_m}{\widehat{\pi}_{\text{c}}}\ \widehat{Y}_{1\cdot}(m)\right] \\ &\quad - \left[\frac{\widehat{e}_{\text{c}}(f)\,\widehat{p}_f}{\widehat{\pi}_{\text{c}}}\ \widehat{Y}_{0\cdot}(f) + \frac{\widehat{e}_{\text{c}}(m)\,\widehat{p}_m}{\widehat{\pi}_{\text{c}}}\ \widehat{Y}_{0\cdot}(m)\right] \\ &= \frac{\widehat{e}_{\text{c}}(f)\,\widehat{p}_f}{\widehat{\pi}_{\text{c}}}\ \widehat{\text{ITT}}(f) + \frac{\widehat{e}_{\text{c}}(m)\,\widehat{p}_m}{\widehat{\pi}_{\text{c}}}\,\widehat{\text{ITT}}(m),\end{aligned}$$

where $\widehat{\text{ITT}}(x)$ denotes the ITT estimate for the subgroup with $X_i = x$. Again, observed program participation (i.e., $D^{\text{obs}}$) is irrelevant under strong PI; the only role stratum membership plays is in subgroup weights.

Finally, we can contrast $\widehat{\text{CACE}}^{\text{Strong PI}}$ with the standard instrumental variable (IV) estimator assuming the exclusion restriction for Never Takers,

$$\widehat{\text{CACE}}^{\text{IV}} = \frac{\widehat{\text{ITT}}}{\widehat{\pi}_c}.$$

The only difference is in the numerator: The strong PI estimator replaces the overall ITT estimate with a weighted average of two subgroup ITT estimates.

### *5.2. Estimation With a General Covariate*

We now extend this setup to the general case of arbitrary, multidimensional covariates. For illustration, we focus on estimating the average outcome for Compliers assigned to control, $\mu_{c0}$, in the one-sided noncompliance setting; other estimates follow similarly (see Ding & Lu, 2017, for a more technical presentation).

For generic $X$, the overall stratum-specific mean, $\mu_{c0}$, is a weighted average across an infinite number of possible subgroups defined by $X_i = x$ (i.e., an integral):

$$\mu_{c0} = \mathbb{E}[\mathbb{E}[Y_i^{\text{obs}}|S_i = c, Z_i = 0, X_i = x]|S_i = c] = \int_x \mu_{c0}(x)\,p(x|c)\,dx.$$

This is the law of iterated expectations conditional on stratum membership, where randomization allows us to drop the conditioning on $Z$. As in the binary case above, we can use Bayes's rule to replace $p(x|c)$, which yields

$$\mu_{c0} = \int_x \mu_{c0}(x)\, p(x|c)\, dx = \int_x \mu_{c0}(x) \frac{e_c(x)\, p(x)}{\pi_c}\, dx. \quad (12)$$

We can calculate the empirical analog of this integral via a summation over all units assigned to the control group, plugging in $Y_i^{\text{obs}}$ for $\mu_{c0}(x_i)$ and $\widehat{e}_c(x_i)$ for $e_c(x)$.[3] The empirical estimate of $p(x)$ weights all units equally in the sum, which introduces a $1/N_0$ term. In addition, $\widehat{\pi}_c$ is the average of $\widehat{e}_c(x_i)$ over these units, $\widehat{\pi}_c = \sum_i \widehat{e}_c(x_i)/N_0$. This yields

$$\widehat{\mu}_{c0} = \frac{1}{N_0}\sum_i^{N_0} \frac{Y_i^{\text{obs}} \cdot \widehat{e}_c(x_i)}{\widehat{\pi}_c} = \frac{\sum_i^{N_0} Y_i^{\text{obs}} \cdot \widehat{e}_c(x_i)}{\sum_i^{N_0} \widehat{e}_c(x_i)} \quad (13)$$

Our estimate of the overall average outcome for Compliers assigned to control is a principal score-weighted average of all individuals in the control group. This reduces to Equation 11 for binary $X$.

### *5.3. Principal Score Weighting*

Importantly, the weights in Equation 13 depend only on the estimated principal score, $\widehat{e}_c(x)$, which is scalar regardless of the number of covariates. Thus, so long as we can estimate $e_{\text{c}}(x)$, we can use Equation 13. This is the principal score weighting method of Stuart and Jo (2015) and Ding and Lu (2017).

*5.3.1. One-sided noncompliance.* Following Equation 13, we estimate $\mu_{c0}$ and $\mu_{n0}$ via a principal score-weighted average of outcomes for individuals assigned to control:

$$\widehat{\mu}_{\text{c}0} = \frac{\sum_i^{N_0} Y_i^{\text{obs}} \cdot \widehat{e}_{\text{c}}(x_i)}{\sum_i^{N_0} \widehat{e}_{\text{c}}(x_i)}, \qquad \widehat{\mu}_{\text{n}0} = \frac{\sum_i^{N_0} Y_i^{\text{obs}} \cdot \widehat{e}_{\text{n}}(x_i)}{\sum_i^{N_0} \widehat{e}_{\text{n}}(x_i)}.$$

As in the binary covariate case, there are two possible estimators for $\mu_{c1}$ and $\mu_{n1}$. First, under weak PI, we can directly estimate $\widehat{\mu}_{c1}^{\text{Weak PI}} = \widehat{Y}_{11}$ and $\widehat{\mu}_{n1}^{\text{Weak PI}} = \widehat{Y}_{10}$. Second, under strong PI, we can estimate these quantities via a principal score-weighted average of individuals assigned to treatment:

$$\widehat{\mu}_{c1}^{\text{Strong PI}} = \frac{\sum_i^{N_1} Y_i^{\text{obs}} \cdot \widehat{e}_c(x_i)}{\sum_i^{N_1} \widehat{e}_c(x_i)}, \qquad \widehat{\mu}_{n1}^{\text{Strong PI}} = \frac{\sum_i^{N_1} Y_i^{\text{obs}} \cdot \widehat{e}_n(x_i)}{\sum_i^{N_1} \widehat{e}_n(x_i)}.$$

Again, we can estimate the CACE and NACE by subtracting the estimated means under treatment and control:

$$\widehat{\text{CACE}}^{\text{PI}} = \widehat{\mu}_{c1}^{\text{PI}} - \widehat{\mu}_{c0} \qquad \widehat{\text{NACE}}^{\text{PI}} = \widehat{\mu}_{n1}^{\text{PI}} - \widehat{\mu}_{n0},$$

where the superscript PI denotes either weak or strong PI.

This is the weighting estimator proposed by Stuart and Jo (2015), with extensions in Ding and Lu (2017). A minor complication is how each paper normalizes the weights; we follow Ding and Lu (2017), who normalize weights separately for treatment and control groups. Note that, while Jo and Stuart (2009) and Stuart and Jo (2015) formally assume strong PI, their proposed weighting estimator is nonetheless valid under weak PI.

*5.3.2. Two-sided noncompliance.* For two-sided noncompliance, we consider three sets of assumptions: (1) strong PI, (2) weak PI for all compliance types, and (3) weak PI for Always Takers and Compliers with an exclusion restriction for Never Takers. This last set of assumptions is particularly important because we believe it is the most plausible of the three for our running example of HSIS.

*Average outcomes for Compliers.* We begin with estimating the average outcomes for Compliers. For $\mu_{c0}$:

$$\widehat{\mu}_{c0}^{\text{Strong PI}} = \frac{\sum_i Y_i^{\text{obs}} \cdot \widehat{e}_c(x_i)}{\sum_i \widehat{e}_c(x_i)} \qquad \text{for } i : Z_i = 0,$$

$$\widehat{\mu}_{c0}^{\text{Weak PI}} = \frac{\sum_i Y_i^{\text{obs}} \cdot \frac{\widehat{e}_c(x_i)}{\widehat{e}_c(x_i) + \widehat{e}_n(x_i)}}{\sum_i \frac{\widehat{e}_c(x_i)}{\widehat{e}_c(x_i) + \widehat{e}_n(x_i)}} \qquad \text{for } i : Z_i = 0, D_i^{\text{obs}} = 0,$$

$$\widehat{\mu}_{c0}^{\text{PI, ER}} = \frac{\widehat{\pi}_c + \widehat{\pi}_n}{\widehat{\pi}_c} \widehat{Y}_{00} - \frac{\widehat{\pi}_n}{\widehat{\pi}_c} \widehat{Y}_{10}.$$

The estimator under strong PI is the principal score-weighted average over all individuals with $Z_i = 0$, ignoring observed behavior. The estimator under weak PI is the principal score-weighted average only over individuals with $Z_i = 0$ and $D_i = 0$, with weights normalized to exclude the probability of being an Always Taker; that is, we weight by

$$\widehat{\mathbb{P}}\{S_i = \text{c} | \mathbf{X}_i = \mathbf{x}, S_i \in \{c, n\}\} = \frac{\widehat{e}_c(x_i)}{\widehat{e}_c(x_i) + \widehat{e}_n(x_i)}.$$

The estimator under the exclusion restriction for the Never Takers is the standard IV estimate for $\mu_{c0}$.

For $\mu_{c1}$, we have

$$\widehat{\mu}_{c1}^{\text{Strong PI}} = \frac{\sum_i Y_i^{\text{obs}} \cdot \widehat{e}_c(x_i)}{\sum_i \widehat{e}_c(x_i)} \quad \text{for } i : Z_i = 1,$$

$$\widehat{\mu}_{c1}^{\text{Weak PI}} = \frac{\sum_i Y_i^{\text{obs}} \cdot \frac{\widehat{e}_c(x_i)}{\widehat{e}_c(x_i) + \widehat{e}_a(x_i)}}{\sum_i \frac{\widehat{e}_c(x_i)}{\widehat{e}_c(x_i) + \widehat{e}_a(x_i)}} \quad \text{for } i : Z_i = 1, D_i^{\text{obs}} = 1,$$

$$\widehat{\mu}_{c1}^{\text{PI,ER}} = \widehat{\mu}_{c1}^{\text{Weak PI}}.$$

Again, the estimator under strong PI includes all individuals assigned to treatment; the estimator under weak PI only includes those with $D_i^{\text{obs}} = 1$.

*Average outcomes for Always Takers.* For $\mu_{a0}$:

$$\widehat{\mu}_{a0}^{\text{Strong PI}} = \frac{\sum_i Y_i^{\text{obs}} \cdot \widehat{e}_a(x_i)}{\sum_i \widehat{e}_a(x_i)} \quad \text{for } i : Z_i = 0,$$

$$\widehat{\mu}_{a0}^{\text{Weak PI}} = \widehat{Y}_{01}.$$

The weak PI estimate is straightforward because we observe Always Takers assigned to control. For $\mu_{a1}$:

$$\widehat{\mu}_{a1}^{\text{Strong PI}} = \frac{\sum_i Y_i^{\text{obs}} \cdot \widehat{e}_a(x_i)}{\sum_i \widehat{e}_a(x_i)} \quad \text{for } i : Z_i = 1$$

$$\widehat{\mu}_{a1}^{\text{Weak PI}} = \frac{\sum_i Y_i^{\text{obs}} \cdot \frac{\widehat{e}_a(x_i)}{\widehat{e}_c(x_i) + \widehat{e}_a(x_i)}}{\sum_i \frac{\widehat{e}_a(x_i)}{\widehat{e}_c(x_i) + \widehat{e}_a(x_i)}} \quad \text{for } i : Z_i = 1, D_i^{\text{obs}} = 1.$$

As above, the strong and weak PI estimates differ by whether they are restricted to those who are observed to participate. For both $\mu_{a0}$ and $\mu_{a1}$, the estimates are the same regardless of whether we assume the exclusion restriction for Never Takers; thus, $\widehat{\mu}_{az}^{\text{PI, ER}} = \widehat{\mu}_{az}^{\text{Weak PI}}$ for $z = 0, 1$.

*Average outcomes for Never Takers.* For $\mu_{n0}$:

$$\widehat{\mu}_{n0}^{\text{Strong PI}} = \frac{\sum_i Y_i^{\text{obs}} \cdot \widehat{e}_n(x_i)}{\sum_i \widehat{e}_n(x_i)} \quad \text{for } i : Z_i = 0,$$

$$\widehat{\mu}_{n0}^{\text{Weak PI}} = \frac{\sum_i Y_i^{\text{obs}} \cdot \frac{\widehat{e}_n(x_i)}{\widehat{e}_c(x_i) + \widehat{e}_n(x_i)}}{\sum_i \frac{\widehat{e}_n(x_i)}{\widehat{e}_c(x_i) + \widehat{e}_n(x_i)}} \quad \text{for } i : Z_i = 0, D_i^{\text{obs}} = 0,$$

$$\widehat{\mu}_{n0}^{\text{PI, ER}} = \widehat{Y}_{10}.$$

Again, the strong and weak PI estimates differ by whether they are restricted to the group with $D_i^{\text{obs}} = 0$. The exclusion restriction states that $\mu_{n0} = \mu_{n1}$;

thus, $\widehat{\mu}_{n0}^{\text{PI, ER}} = \widehat{Y}_{10}$, the average outcome for the observed Never Takers assigned to treatment.

Finally, for $\mu_{n1}$:

$$\widehat{\mu}_{n1}^{\text{Strong PI}} = \frac{\sum_i Y_i^{\text{obs}} \cdot \widehat{e}_n(x_i)}{\sum_i \widehat{e}_n(x_i)} \quad \text{for } i : Z_i = 1,$$

$$\widehat{\mu}_{n1}^{\text{Weak PI}} = \widehat{Y}_{10},$$

$$\widehat{\mu}_{n1}^{\text{PI,ER}} = \widehat{\mu}_{n1}^{\text{Weak PI}}.$$

Recall that, under the exclusion restriction, NACE $= 0$.

*5.3.3. Quantifying uncertainty in impact estimates.* We can use standard results from causal inference to calculate the variance of this estimator under each set of assumptions. Assuming that the principal score is known, we have

$$\text{var}(\widehat{\text{CACE}}) = \text{var}(\widehat{\mu}_{c1}) + \text{var}(\widehat{\mu}_{c0}),$$

where $\text{var}(\widehat{\mu}_{cz})$ is computed via the variance of the weighted mean and where we ignore the covariance between the two terms under the superpopulation perspective (see, e.g., Imbens & Rubin, 2015).

This formulation highlights the precision gains from stronger identifying assumptions. For example, consider the variance of the two estimators for $\mu_{c1}$ in the one-sided noncompliance setting (i.e., no Always Takers) where, to give intuition, we condition the variance on the sample covariates:

$$\text{var}\left(\widehat{\mu}_{c1}^{\text{Weak PI}}\right) = \text{var}\left(\frac{\sum_i^{N_1} Y_i^{\text{obs}} \cdot 1_{\{D_i=1\}}}{\sum_i^{N_1} 1_{\{D_i=1\}}}\right) = \frac{\text{var}(Y_i(1)|S_i = c)}{n_{11}},$$

$$\text{var}\left(\widehat{\mu}_{c1}^{\text{Strong PI}}\right) = \text{var}\left(\frac{\sum_i^{N_1} Y_i^{\text{obs}} \cdot e_c(x_i)}{\sum_i^{N_1} e_c(x_i)}\right) = \frac{\sum_i^{N_1} e_c(x_i)^2}{\sum_i^{N_1} e_c(x_i)} \cdot \frac{\text{var}(Y_i(1)|S_i = c)}{n_{11}},$$

where $n_{11} = \sum_i^{N_1} e_c(x_i) = \sum_i^{N_1} 1_{\{D_i=1\}}$. Since $e_c(x_i)$ are probabilities between 0 and 1, $\sum_i^{N_1} e_c(x_i)^2 \leq \sum_i^{N_1} e_c(x_i)$. Thus, the theoretical variance under strong PI will be smaller than the corresponding variance under weak PI, except in the special case of perfect compliance (i.e., $e_c(x_i) = 1$ for all $i$). At the same time, we worry that the estimator assuming strong PI might be biased if the underlying, very strong assumptions do not hold. Thus, researchers must balance these considerations in deciding on the particular estimator.

In practice, we want to incorporate uncertainty from principal score estimation as well as from sampling uncertainty. The standard case-resampling bootstrap is a natural way to account for both stages in the principal score estimation procedure. In particular, the researcher first obtains a point estimate by running the entire procedure on the original data set. Then, the researcher generates $B$ (typically 1,000 or more) *bootstrap* data sets by sampling $N$ rows with replacement from the original data set. Finally, the researcher runs both steps of the principal score estimation procedure, both fitting the principal score and estimating the impacts, separately for each of the $B$ bootstrap data sets. The 2.5th and 97.5th quantiles of the bootstrap distribution therefore generate a 95% confidence interval. See Aronow and Carnegie (2013) and Ding and Lu (2017) for additional discussion of the bootstrap in related settings.

### 5.4. Assessing Principal Score Fit

Just as with propensity scores, a key concern is whether the principal score model has been correctly specified. Ding and Lu (2017) offer one promising approach for balance checks in this setting. Due to randomization, the distribution of covariates should be the same for individuals assigned to treatment and control within each principal stratum. The challenge is that we do not observe principal strata directly and therefore cannot easily compare these distributions. The key insight is that we can use the weighting scheme for estimating stratum-specific average outcomes to estimate stratum-specific covariate distributions. Specifically, Ding and Lu (2017) show that we can estimate any stratum-specific function of covariates, $h(x)$, via a principal score weighted average. For example, in the case of one-sided noncompliance, we can estimate $h(x)$ for Compliers assigned to control via the weighted average among all control individuals:

$$\widehat{h(x)}_{c0} = \frac{\sum_i h(x_i) \cdot \widehat{e}_c(x_i)}{\sum_i \widehat{e}_c(x_i)} \qquad \text{for } i : Z_i = 0.$$

Continuing this example, we can directly observe, say, the average age of Compliers assigned to treatment, $\overline{\text{age}}_{c1}$. We can then substitute $h(x_i) = \text{age}_i$ into the above formula to estimate $\overline{\text{age}}_{c0}$, which should equal $\overline{\text{age}}_{c1}$ in expectation. A mismatch between these means suggests a poorly estimated principal score model.

As with standard propensity score methods, we can try to improve balance by refitting the model with additional terms, such as interactions and higher order polynomials (e.g., Imbens & Rubin, 2015), repeating the procedure until we no longer see imbalance. More generally, we can assess covariate balance via a range of metrics. We focus on the *normalized difference* for each covariate within each principal stratum (see Imbens & Rubin, 2015),

$$\widehat{\Delta}_s = \frac{\overline{X}_{s1} - \overline{X}_{s0}}{\sqrt{(s_{s1}^2 + s_{s0}^2)/2}},$$

where the covariate mean and standard deviation, $\overline{X}_{sz}$ and $s^2_{sz}$, are either calculated directly from the observed data or via this weighting method described above. In the observational study setting, normalized differences greater than 0.1 are often thought to require additional adjustment, such as via linear regression, and suggest poor covariate balance. See, for example, Imbens and Rubin (2015). We return to this point in Section 7.

This logic extends to two-sided noncompliance. Specifically, we can compare the estimated Never Taker covariate distribution under control with the observed distribution under treatment. Similarly, we can compare the estimated Always Taker covariate distribution under control with the observed distribution under treatment. For Compliers, we must compare the weighted estimates from each treatment arm.

## 6. Applications

### *6.1. Application to JOBS II*

We now replicate the principal score analysis of JOBS II from Stuart and Jo (2015), continuing the discussion from Section 2 (see also Jo & Stuart, 2009). Following their analysis, we focus on the $N = 410$ high-risk, unemployed workers. Of this group, 273 were randomly encouraged to attend the job training ($Z_i = 1$) and 137 were not ($Z_i = 0$). Among those encouraged to attend, only 160, or 59%, actually attended at least one of the five job training sessions ($D_i^{\text{obs}} = 1$). The outcome of interest is a depression score measured 6 months after baseline (here we standardize the outcome by the mean and variance of the control group). The overall ITT for this standardized outcome is $-0.29$, which indicates that the encouragement to treatment decreased depression symptoms. The goal of this analysis is to estimate the impact of randomization for the subgroups of Compliers and Never Takers. In particular, there is a concern that the exclusion restriction for Never Takers might not hold here; that is, $\text{NACE} \neq 0$. See Jo and Stuart (2009) for additional discussion.

Following Stuart and Jo (2015), we focus on seven covariates measured at baseline: depression, motivation, sense of mastery, economic hardship, age, gender, and years of schooling. We estimate the principal score via a logistic regression of $D$ on $\mathbf{X}$ among those individuals assigned to treatment. We then assess the fit of the principal score model using the procedure in Section 5.4. Figure 1 shows the normalized covariate mean differences for the overall data set as well as for Compliers and Never Takers.

First, even before we turn to principal strata, there are meaningful covariate imbalances across randomization groups in this experimental subset: The encouragement group has lower depression levels but greater economic hardship at baseline than the control group. It is therefore unsurprising that we see similar (estimated) imbalances among Compliers and Never Takers. In addition, there

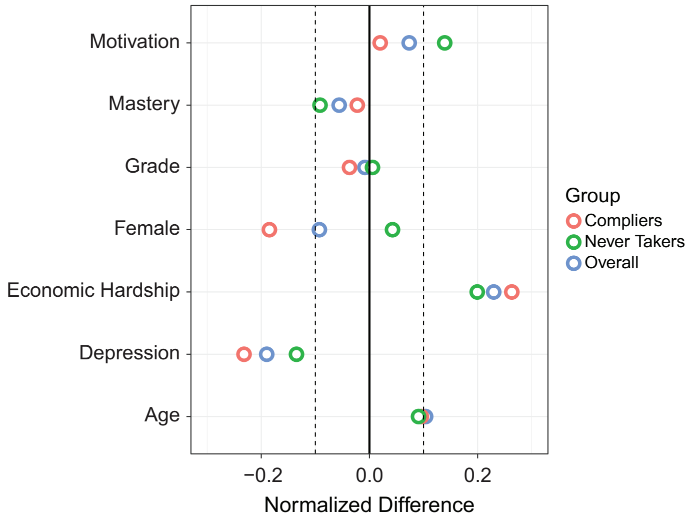


FIGURE 1. *Normalized treatment to control mean differences for seven key covariates in the JOBS II evaluation.*

seems to be meaningful imbalance on proportion female: 53% among Compliers assigned to encouragement relative to 62% among those assigned to control. Taken together, these imbalances suggest that the simple logistic regression with main effects is not a sufficiently rich principal score model. Unfortunately, more complex models (including adding higher order terms, interactions, and similar modifications) do not improve these imbalances, indicating that the imbalances across randomization groups are too large to correct via principal score modeling alone. We therefore interpret the resulting principal causal effects with caution.

Next, we use the principal score to estimate the stratum means. The first three panels of Figure 2 show the means and bootstrapped 95% confidence intervals separately by observed group under three different assumptions: (1) strong PI, (2) weak PI, and (3) the exclusion restriction for Never Takers. The bottom-right panel shows the impact by stratum, which is the difference in means from the other panels.

There are several key takeaways. First, the means are remarkably stable across the three assumptions. Second, there is no evidence to reject strong PI since $\widehat{\mu}_{c1}$ is essentially unchanged under strong versus weak PI. Finally, the estimated impact for Never Takers is quite close to zero under both strong and weak PI, which is consistent with the exclusion restriction.

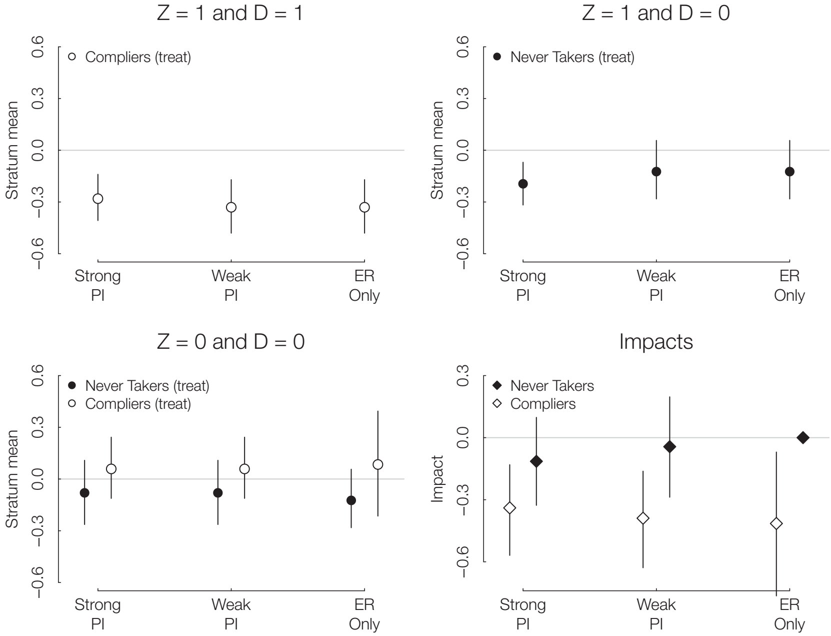


FIGURE 2. *Means and impacts by principal stratum under different assumptions, with bootstrap 95% confidence intervals.*

### *6.2. Application to HSIS*

We now turn to HSIS. In roughly 350 Head Start centers, the offer of enrollment was randomly assigned among eligible children. Although the original study collected a range of outcomes, we focus on the Peabody Picture Vocabulary Test (PPVT), a widely used measure of cognitive ability in early childhood. For this outcome, the ITT estimate was 0.14 effect size units (i.e., in terms of the control group standard deviation), which was modest relative to other estimates of Head Start's impact (see, e.g., Bloom & Weiland, 2014). For interpretability, we have centered the outcome relative to the overall control group mean in the sample.

Noncompliance in HSIS was meaningful. Of those offered a spot, 18% of children in our analysis sample were Never Takers who did not actually enroll (i.e., $\widehat{\pi}_n = 0.18$). In addition, 13% of children not offered the opportunity to enroll were Always Takers who nonetheless enrolled in a Head Start center during the study period (i.e., $\widehat{\pi}_a = 0.13$). Roughly half of the observed Always Takers enrolled in the center of randomization (i.e., where they were formally denied access to the program for that year) and half enrolled in a different Head Start Center (Puma et al., 2010). Finally, this leaves $\widehat{\pi}_c = 0.69$ for Compliers in the sample.

Since the goal of the study is to estimate the effect of enrolling in Head Start, the standard approach would be to invoke the usual instrumental variables assumptions to estimate the CACE: monotonicity and the exclusion restrictions for Always Takers and Never Takers (Angrist, Imbens, & Rubin, 1996). Although both monotonicity and the exclusion restriction for Never Takers are highly plausible in this case, the exclusion restriction for Always Takers is somewhat more controversial. In particular, as Gibbs, Ludwig, and Miller (2011) argue, centers of enrollment for Always Takers could systematically differ from their centers of randomization (see also Bloom & Weiland, 2014). Thus, we propose using principal score methods to explore the effect of the exclusion restriction for Always Takers on estimates of the CACE. Following earlier analyses (Ding, Feller, & Miratrix, 2016) and to simplify exposition, we restrict our attention to a complete-case subset of HSIS, with $N_1 = 2{,}238$ in the treatment group and $N_0 = 1{,}348$ in the control group.[4] For covariates, we will adopt the rich set of child- and family-level covariates used in the original HSIS analysis of Puma, Bell, Cook, Heid, and Shapiro (2010), including pretest score, child's age, child's race, mother's education level, and mother's marital status. In total, there are $k = 20$ covariates after recoding factor variables. Despite these important covariates, PI assumptions are nonetheless quite heroic in this context.

First, we fit principal score models using the "marginal method" in Section 5.3. That is, we estimate two separate logistic regressions by treatment arm to estimate $\widehat{e}_a(x_i)$ and $\widehat{e}_n(x_i)$ and then subtract to estimate $\widehat{e}_c(x_i) = 1 - \widehat{e}_a(x_i) - \widehat{e}_n(x_i)$. In this example, we use only main effects; adding in higher order interactions gave comparable covariate balance. We then assess covariate balance given the estimated principal score via the normalized difference for each child-level covariate within each principal stratum, as described in Section 5.4. All differences are below 0.1 in absolute value, suggesting that there is good covariate balance given the principal score. We also estimated the principal score via the "joint method," which restricts the estimated principal scores to be between 0 and 1; this yielded nearly identical results.

We then estimated principal causal effects under our different assumptions to see how our estimates changed. Figures 3 and 4 show the estimated principal stratum means and impacts, respectively, given (1) strong PI, (2) weak PI, (3) weak PI plus the exclusion restriction for the Never Takers, and (4) exclusion restrictions for both Always Takers and Never Takers (i.e., standard IV). We obtain 95% confidence intervals via a standard case-resampling bootstrap.

We start by assessing the exclusion restriction for Never Takers. As shown in the bottom-right panel of Figure 3, the estimate for $\mu_{n0}$ changes very little with and without the exclusion restriction. Figure 4 shows the same change in terms of impacts, which emphasizes that the estimate for NACE under weak PI is not meaningfully different from zero (i.e., the exclusion restriction). Thus, estimates assuming weak PI for Compliers and Never Takers do not yield any evidence against the exclusion restriction for Never Takers.

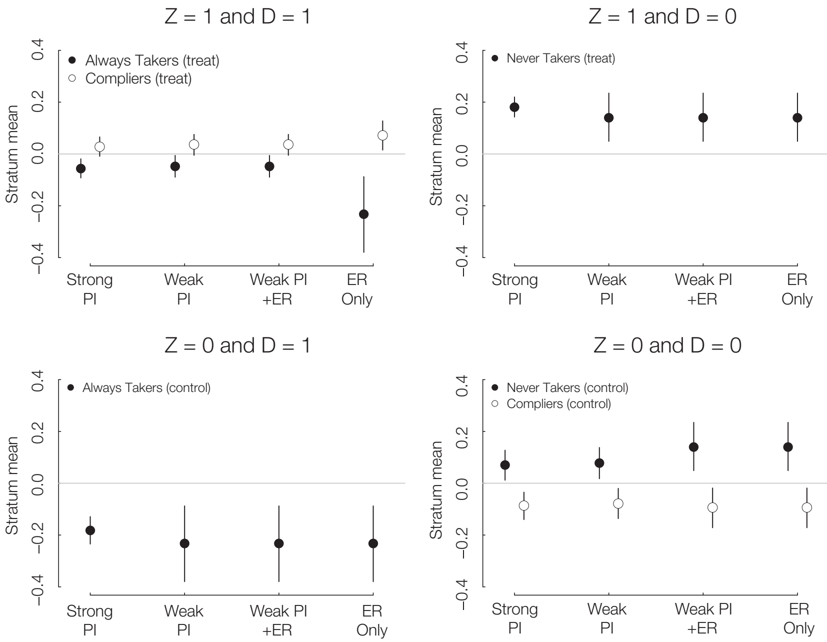


FIGURE 3. *Outcome means by principal stratum under different assumptions, with bootstrap 95% confidence intervals. "Weak PI + ER" refers to weak PI for Compliers and Always Takers plus the exclusion restriction for Never Takers. "ER only" refers to the exclusion restriction for both Always Takers and Never Takers (i.e., the standard instrumental variables assumptions).*

By contrast, consider the exclusion restriction for Always Takers. As shown in the top-left panel of Figure 3, the estimate for $\mu_{a1}$ under weak PI is quite different from under an exclusion restriction. Figure 4 displays the same change in terms of impacts. Although estimates for the AACE are highly uncertain, they are nonetheless consistently positive and away from zero. This result suggests that, based on observable characteristics alone, we should be wary of the exclusion restriction for Always Takers in HSIS. In the end, however, the estimates for the CACE across these different assumptions are quite similar, as shown in Figure 4.

We also considered the testable implications of strong PI in this example. The top-right and bottom-left panels of Figure 3 show the estimates for $\mu_{a0}$ and $\mu_{n1}$, respectively, the two principal stratum means that we can directly estimate in this example. Since estimates are largely unchanged under strong PI and weak PI, we do not find evidence against strong PI here. Nevertheless, since weak PI is the strictly weaker assumption, we would typically prefer that in practice, even though there is a meaningful loss in precision.

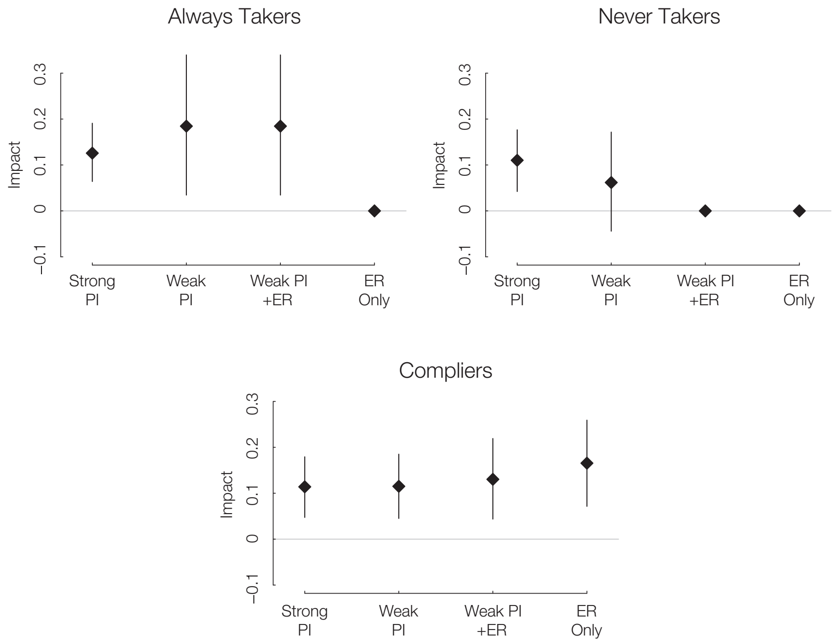


FIGURE 4. *Impacts by principal stratum under different assumptions, with bootstrap 95% confidence intervals. "Weak PI + ER" refers to weak PI for Compliers and Always Takers plus the exclusion restriction for Never Takers. "ER only" refers to the exclusion restriction for both Always Takers and Never Takers (i.e., the standard instrumental variables assumptions).*

## 7. Discussion

Although principal score methods are gaining popularity in the social sciences, these methods remain poorly understood. We make two important contributions in this article. First, we review the literature on principal score methods and clarify the assumptions necessary to use these approaches in practice. Second, we demonstrate how researchers can mix and match PI with other assumptions, such as the exclusion restriction, in more complex examples. We use this development to assess the impact of randomization on Always Takers in HSIS, finding evidence that there is in fact a positive treatment effect for this group, at least under weak PI. We also assess the impact of randomization on Never Takers in the JOBS II evaluation, finding no evidence against the exclusion restriction in this case.

Overall, we argue that the assumptions underlying principal score analysis are typically unrealistic. Thus, researchers should be wary of relying solely on these approaches. At the same time, we believe that methods based on principal scores can be useful as one of many ways to estimate the same principal causal effects. For example, researchers using model-based approaches should consider estimating the same quantities via principal score methods; if the estimates diverge, results may be particularly sensitive to modeling choices or other assumptions. More broadly, principal score methods can be useful as simple, exploratory tools.

We briefly discuss several directions for future work. First, while we only explore noncompliance here, extending these results to more complex settings is straightforward. See Page et al. (2015) for some possible settings relevant to education and behavioral science. Second, while we used simple estimators in the main text, principal score models can be quite rich. Feller (2015), for example, estimates the principal score for HSIS using a Bayesian hierarchical model that accounts for the multilevel structure in the experiment. We could also consider nonparametric or machine learning methods, which have proved effective with propensity scores (e.g., B. K. Lee, Lessler, & Stuart, 2010).

Third, our discussion has focused on the use of covariates solely for justifying PI and estimating the principal score. In practice, we can also leverage covariates that are predictive of the outcome to sharpen inference for the causal effects themselves. Jo and Stuart (2009) propose a straightforward strategy of a weighted regression of $Y^{\text{obs}}$ on $Z$ and $X$ with the relevant principal score weights. Ding and Lu (2017) borrow methods from survey sampling and discuss model-assisted estimation, which reduces to the strategy in Jo and Stuart (2009) in certain settings. How best to incorporate covariates is a promising question for additional study (see also Mealli et al., 2016).

Another critical direction for future work is sensitivity analysis. Ding and Lu (2017) take an important step in this direction, proposing formal sensitivity analyses analogous to those approaches for observational studies first introduced by Rosenbaum and Rubin (1983). This is especially important because, as discussed above, PI assumptions are quite strong. One potentially fruitful approach—essentially a quick-and-dirty sensitivity analysis—is to compare principal score estimates and their corresponding nonparametric bounds. The bounds give a range of plausible parameter values, and the principal score estimate gives a "reasonable guess" within this interval as to where the truth might be. Furthermore, a principal score estimate outside these bounds—though unlikely to occur in practice—would be strong evidence against PI. This is an attractive approach given available tools.

Finally, principal scores are useful objects for describing trends in data even in the absence of PI assumptions, just as the propensity score can be useful in settings other than observational studies. In particular, they can be used to describe trends in how individuals respond to the offer of treatment, which is often of substantive interest in its own right. We are also currently exploring how, even without the ignorability assumptions, principal scores can be used to tighten nonparametric bounds (see also Long & Hudgens, 2013). We anticipate that there will be many other uses.

## Appendix A

### A.1. Discrete Subgroup Method

Schochet and Burghardt (2007) suggest estimating the CACE and NACE via discrete subgroups. First, let $\pi_c \equiv \mathbb{P}\{S_i = c\}$ be the overall proportion of Compliers in the population, and let $C_i$ be an indicator for whether individual $i$ is a Complier. Next, define $\widehat{C}_i = \mathbb{I}\{\widehat{e}_c(x_i) \geq \widehat{\pi}_c\}$, as the indicator for whether individual $i$ is predicted to be a Complier based on being above a given threshold. Finally estimate the CACE via the ITT for those individuals with $\widehat{C}_i = 1$ and estimate the NACE via the ITT for those individuals with $\widehat{C}_i = 0$.

The intuition is that the predictive model does not depend on outcomes or treatment assignment. The identified subgroups, which we might call "likely Compliers" and "likely Never Takers," are therefore pretreatment subgroups and can be described and explored just as any other pretreatment subgroup. Having such easily interpretable groups and being able to estimate them in a straightforward way is appealing.

To illustrate this approach, we return to the simple case with binary $X$. First, without loss of generality, assume that individuals with $X_i = f$ are more likely to be Compliers than those with $X_i = m$, i.e., $e_c(f) > e_c(m)$. Under strong PI, we can estimate the CACE via the weighted average of the subgroup ITT effects for females and males. So long as there is sufficient imbalance between females and males, the discrete subgroup method will only use the first term of this equation to estimate $\mathrm{ITT}_c$:

$$\widehat{\mathrm{CACE}}^{\mathrm{Sub}} = \widehat{\mathrm{ITT}}(f).$$

That is, we use the estimated ITT among females as a proxy for the CACE. This only matches the plug-in estimator if $X$ is perfectly predictive of $C$ (i.e., $e_c(f) = 1$), or if there is no impact variation across principal strata (i.e., ITT $=$ CACE $=$ NACE), neither of which is an interesting case. Although it might be possible to motivate this estimator with a different set of assumptions, these are not immediately apparent. These quantities are, however, valid estimates for alternate estimands: the average effects for groups defined by predicted membership. This might be of interest in some settings and, at the very least, is a useful exploratory tool.

### A.2. Comparison With Sequential Ignorability

We offer a brief comparison of PI and sequential ignorability (for additional discussion, see, e.g., VanderWeele, 2011). Sequential ignorability is typically associated with mediation methods and allows the researcher to estimate effects for the entire population. Although appealing, this comes at a cost: We must be able to imagine a hypothetical experiment in which $D$ could plausibly be assigned at random.

The sequential ignorability assumption conceives of *both* $Z$ and $D$ as if they could be randomly assigned, as in a two-stage randomization scheme or factorial design. Under this formulation, we doubly index the potential outcomes as $Y_i(z, d)$, leading to four possible combinations: $Y_i(1, 1)$, $Y_i(1, 0)$, $Y_i(0, 1)$, and $Y_i(0, 0)$. We then state the sequential ignorability assumption as

$$Y_i(z, d) \perp\!\!\!\perp D_i | \mathbf{X}_i, Z_i \text{ for } z \in \{0, 1\} \textit{ and } d \in \{0, 1\}.$$

The analogous estimands to CACE and NACE are therefore

$$\mathrm{ITT}_c^{\mathrm{SI}} = \mathbb{E}\{Y_i(1, 1) - Y_i(0, 0)\},$$

$$\mathrm{ITT}_n^{\mathrm{SI}} = \mathbb{E}\{Y_i(1, 0) - Y_i(0, 0)\}.$$

As in the typical ignorability case, this estimand is defined for the entire (super) population of individuals. By contrast, PI focuses on estimands for specific principal strata. Thus, PI is a more "local" assumption than sequential ignorability.

### A.3. Proofs

This proof is nearly identical to the analogous proofs for the propensity score in Imbens and Rubin (2015). Following that example, we first show that the principal score is indeed a balancing score. For convenience, let $C_i$ be an indicator for whether individual $i$ is a Complier. We wish to show that

$$C_i \perp\!\!\!\perp \mathbf{X}_i | e_c(\mathbf{x}_i),$$

or equivalently,

$$\mathbb{P}\{C_i = 1 | \mathbf{X}_i, e_c(x_i)\} = \mathbb{P}\{C_i = 1 | e_c(x_i)\}.$$

We will show that both sides of the equation equal $e_c(x_i)$. For the left-hand side, $\mathbb{P}\{C_i = 1 | \mathbf{X}_i, e_c(x_i)\} = \mathbb{P}\{C_i = 1 | \mathbf{X}_i\} = e_c(x_i)$. For the right-hand side:

$$\mathbb{P}\{C_i = 1 | e_c(x_i)\} = \mathbb{E}\{C_i | e_c(x_i)\} = \mathbb{E}\{\mathbb{E}\{C_i | \mathbf{X}_i, e_c(x_i)\} | e_c(x_i)\} = \mathbb{E}\{e_c(x_i) | e_c(x_i)\} = e_c(x_i).$$

Therefore, the principal score is a balancing score.

Second, we show that if strong PI holds given $\mathbf{X}_i$, strong PI also holds given $e_c(x_i)$. We show this for $Y_i(0)$, with an identical argument for $Y_i(1)$. To do this, we need to show that

$$Y_i(0) \perp\!\!\!\perp C_i | e_c(x_i),$$

holds, or equivalently,

$$\mathbb{P}\{C_i = 1 | Y_i(0), e_c(x_i)\} = \mathbb{P}\{C_i = 1 | e_c(x_i)\}.$$

To show this:

$$\begin{aligned} \mathbb{P}\{C_i = 1 | Y_i(0), e_c(x_i)\} &= \mathbb{E}\{C_i | Y_i(0), e_c(x_i)\} = \mathbb{E}\{\mathbb{E}\{C_i | Y_i(0), \mathbf{X}_i, e_c(x_i)\} | Y_i(0), e_c(x_i)\} \\ &= \mathbb{E}\{\mathbb{E}\{C_i | e_c(x_i)\} | Y_i(0), e_c(x_i)\} = \mathbb{E}\{C_i | e_c(x_i)\} \\ &= \mathbb{P}\{C_i = 1 | e_c(x_i)\}, \end{aligned}$$

where we use PI and the fact that the principal score is a balancing score to go from the second to third lines. Therefore, strong PI also holds, given $e_c(x_i)$. The same argument applies for the various weak PI assumptions.

## Authors' Note

All opinions expressed in the article and any errors that it might contain are solely the responsibility of the authors.


## Acknowledgment

We thank Alberto Abadie, Zach Branson, Peng Ding, Jennifer Hill, Jiannan Lu, Don Rubin, Elizabeth Stuart, Lo-Hua Yuan, and members of the Spencer group for helpful comments and discussion as well as seminar participants at the 2015 Society for Research on Educational Effectiveness meeting.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: A.F. and L.M. gratefully acknowledge financial support 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" and from the Institute for Education Science (IES Grant #R305D150040).


## Notes

1. Note that this article focuses on superpopulation estimands, which appear to be the objects of interest in the principal score literature. We are not aware of any discussion of finite sample versus superpopulation estimands in this setting. See Imbens and Rubin (2015) for further discussion of finite versus superpopulation inference.
2. Jo and Stuart (2009) and Ding and Lu (2017), among others, write these assumptions in terms of full stochastic independence, that is, $\perp\!\!\!\perp$. Since the formal identification results only require mean independence (rather than full stochastic independence), we use mean independence throughout for clarity of exposition. We argue that this distinction is fairly unimportant in this setting: Although mean independence is technically weaker than full stochastic independence, it is difficult to imagine a real-world situation in which PI holds in terms of mean independence but not full stochastic independence.
3. In theory, we could use nonparametric regression to estimate these quantities, for example, we could replace $Y_i^{\text{obs}}$ with a regression estimate.
4. Although we do not do so here, it is straightforward to incorporate inverse probability weights under the assumption of missing at random. In particular,

the researcher can construct overall weights by multiplying the (inverse) weights for missingness and the (direct) weights for principal scores.

## References

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

Abadie, A., Chingos, M. M., & West, M. R. (2016). *Endogenous stratification in randomized experiments* (NBER Working Paper), NBER, Cambridge, MA.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455.

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

Bein, E. (2015). *Proxy variable estimators for principal stratification analyses* (Abt Associates Working Paper), Abt Associates, Cambridge, MA.

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

Crépon, B., Devoto, F., Duflo, E., & Parienté, W. (2015). Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco. *American Economic Journal: Applied Economics*, *7*, 123–150.

Ding, P., Feller, A., & Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*, 655–671.

Ding, P., Geng, Z., Yan, W., & Zhou, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *Journal of the American Statistical Association*, *106*, 1578–1591.

Ding, P., & Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*, 757–777.

Feller, A. (2015). *Essays in causal inference and public policy*. Ph.D. thesis, Harvard University, Cambridge.

Feller, A., Greif, E., Miratrix, L., & Pillai, N. (2016). Principal stratification in the Twilight Zone: Weakly separated components in finite mixture models. Arxiv 1602.06595. Retrieved from https://arxiv.org/abs/1602.06595

Follmann, D. A. (2000). On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association*, *95*, 1101–1109.

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

Gibbs, C., Ludwig, J., & Miller, D. L. (2011). Does head start do any lasting good? In Martha J. Bailey & S. Danziger (eds.), *Legacies of the war on poverty* (pp. 39–65). New York, NY: Russell Sage Foundation.

Grilli, L., & Mealli, F. (2008). Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification. *Journal of Educational and Behavioral Statistics*, *33*, 111–130.

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

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

Hsu, J. Y., & Small, D. S. (2014). Discussion on "Dynamic treatment regimes: Technical challenges and applications." *Electronic Journal of Statistics*, *8*, 1301–1308.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, *85*, 765–769.

Imbens, G., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences*. Cambridge, MA: Cambridge University Press.

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

Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics*, *27*, 385–409.

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

Joffe, M. M., Small, D., & Hsu, C.-Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science*, *22*, 74–97.

Joffe, M. M., Ten Have, T. R., & Brensinger, C. (2003). The compliance score as a regressor in randomized trials. *Biostatistics*, *4*, 327–340.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*, 337–346.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, *76*, 1071–1102.

Long, D. M., & Hudgens, M. G. (2013). Sharpening bounds on principal effects with covariates. *Biometrics*, *69*, 812–819.

Mattei, A., Li, F., & Mealli, F. (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*, 2336–2360.

Mealli, F., & Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, *108*, 1120–1131.

Mealli, F., Pacini, B., & Stanghellini, E. (2016). Identification of principal causal effects using additional outcomes in concentration graphs. *Journal of Educational and Behavioral Statistics*, *41*, 463–480.

Miratrix, L., Furey, J., Feller, A., Grindal, T., & Page, L. C. (2017). Bounding, an accessible method for estimating principal causal effects, examined and explained. arXiv 1701.03139. Retrieved from https://arxiv.org/abs/1701.03139

Page, L. C., Feller, A. I., Grindal, T., Miratrix, L., & Somers, M.-A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation*, *36*, 514–531.

Porcher, R., Leyrat, C., Baron, G., Giraudeau, B., & Boutron, I. (2016). Performance of principal scores to estimate the marginal compliers causal effect of an intervention. *Statistics in Medicine*, *35*, 752–767.

Puma, M., Bell, S. H., Cook, R., Heid, C., & Shapiro, G. (2010). *Head start impact study* (Final Report). Washington, DC: Department of Health and Human Services, Administration for Children and Families.

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, *45*, 212–218.

Schochet, P. Z., & Burghardt, J. (2007). Using propensity scoring to estimate program-related subgroup impacts in experimental program evaluations. *Evaluation Review*, *31*, 95–120.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014–4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

Stuart, E. A., & Jo, B. (2015). Assessing the sensitivity of methods for estimating principal causal effects. *Statistical Methods in Medical Research*, *24*, 657–674.

VanderWeele, T. J. (2011). Principal stratification—Uses and limitations. *The International Journal of Biostatistics*, *7*, 1–14.

Zhai, F., Brooks-Gunn, J., & Waldfogel, J. (2014). Head start's impact is contingent on alternative type of care in comparison group. *Developmental Psychology*, *50*, 2572–2586.

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

Zhang, J. L., Rubin, D. B., & Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, *104*, 166–176.

## Authors

AVI FELLER is an assistant professor of Public Policy and Statistics, Goldman School of Public Policy, University of California, Berkeley, CA 94720; email: afeller@berkeley.edu.

FABRIZIA MEALLI is a professor of Department of Statistics, Computer Science, Applications, University of Florence, Florence 50134, Italy; email: mealli@disia.unifi.it.

LUKE MIRATRIX is an assistant professor of Education, Harvard Graduate School of Education, Cambridge, MA 02138; email: luke_miratrix@gse.harvard.edu.

Manuscript received June 8, 2016
First revision received February 9, 2017
Second revision received May 27, 2017
Accepted June 4, 2017