Political Science Research and Methods (2024), **12**, 557–572
doi:10.1017/psrm.2023.38

ORIGINAL ARTICLE

Improving precision through design and analysis in experiments with noncompliance

Erin Hartman[1] and Melody Huang[2]

[1]University of California, Berkeley, CA, US and [2]Harvard University, Cambridge, MA, US
Corresponding author: Erin Hartman; Email: ekhartman@berkeley.edu

(Received 11 July 2022; revised 23 March 2023; accepted 11 May 2023; first published online 1 September 2023)

Abstract
Even in the best-designed experiment, noncompliance can complicate analysis. While the intent-to-treat effect remains identified, randomization alone no longer identifies the complier average causal effect (CACE). Instrumental variables approaches, which rely on the exclusion restriction, can suffer from high variance, particularly when the experiment has a low compliance rate. We provide a framework which broadens the set of design and analysis techniques political science researchers can use when addressing noncompliance. Building on the growing literature about the advantages of ex-ante design decisions to improve precision, we show blocking on variables related to both compliance and the outcome can greatly improve all the estimators we propose. Drawing on work in statistics, we introduce the principal ignorability assumption and a class of principal score weighting estimators, which can exhibit large gains in precision in low compliance settings. We then combine principal ignorability and blocking with a simple estimation strategy to derive a more efficient estimation strategy for the CACE. In a re-evaluation of a study on the effect of GOTV on turnout, we find that the principal ignorability approaches result in confidence intervals roughly half the size of traditional instrumental variable approaches.

Keywords: Experimental design; noncompliance; principal ignorability

1. Introduction

Experimental trials identify causal effects under a relatively minimal set of assumptions. However, while a researcher can control the random assignment to encouragement for a unit to take-up treatment, she often cannot control whether treatment is ultimately received. Identifying the effect among those who receive treatment requires the careful consideration of additional assumptions.

In this paper, we provide a framework for designing and analyzing experiments in the face of noncompliance. We pull together concepts from the experimental design literature, including blocking and placebo-controlled designs, and the literature on identification and estimation under noncompliance with principal stratification. We elucidate the multiple options that political science researchers have for addressing noncompliance, including an approach in the statistical literature, principal ignorability (PI), that is largely overlooked in political science. Finally, we build on the existing literature by incorporating PI into the design stage and proposing a blocked design with a simple blocked-difference-in-means estimator that shows numerous advantages over existing estimation approaches. This framework expands the toolkit of approaches political scientists can use when addressing noncompliance in experiments. A summary of our proposed framework can be found in Figure 5.

To build our framework, we start with a review of principal stratification, the exclusion restriction, and the introduction of PI and principal score weighting (PSW) (Stuart and Jo, 2015; Ding

and Lu, 2017; Feller *et al.*, 2017) to the political science literature. This serves as an alternative to instrumental variables (IV), the most common approach for addressing noncompliance, using a different set of assumptions about subject behavior. The efficiency of the IV estimator decreases exponentially as compliance rates decrease (i.e., they have high variance), motivating our proposal for an alternative, more efficient, approach to addressing noncompliance. The associated principal score methods are largely absent from the current political science literature.[1] These methods provide a promising, flexible way for researchers to analyze experiments with noncompliance and show great potential for improvement in precision with significant noncompliance, when IV estimates are most unstable.

Second we discuss a design consideration, block-randomization, for when noncompliance is an issue. While blocking on prognostic variables is well known to improve precision, we show that blocking on variables related to *both* compliance *and* the outcome prior to treatment assignment can greatly improve precision regardless of which identifying assumption is invoked or estimator is used.

Third, building on the growing literature about the advantages of ex-ante design decisions to improve precision (e.g., Pashley and Miratrix, 2021a, 2021b), this paper introduces a way to combine PI and blocking with a simple estimation strategy for more efficient estimation of the complier average causal effect (CACE). This approach forces balance on important covariates necessary for identification within the experiment and reduces reliance on the modeling assumptions required for PSW.

Our framework suggests blocking on variables related to both compliance and the outcome to improve precision in traditional IV and PI estimators, and it will provide the most benefits when researchers have access to rich covariate information at the design stage; we also show PSW can effectively incorporate such information in the analysis stage. Both the exclusion restriction and principal ignorability are inherently unverifiable, and researchers may be concerned that principal ignorability is unlikely to hold exactly in practice. We provide a set of diagnostic tests and sensitivity analyses that researchers can use to evaluate robustness to violations of PI. Through a series of simulations, we show the potential advantages of PI approaches, including that they dominate IV on a mean-squared error criterion due to their low variance even when principal ignorability fails to hold to some extent and bias may still be a concern, demonstrating the importance of having multiple options for identification and estimation in our framework.

We re-evaluate the Green *et al.* (2003) Get-Out-the-Vote experiment, a multi-site voter mobilization experiment on the effects of personal canvassing on verified voter turnout, where compliance was roughly 29%. We find that the PI approaches for estimating the CACE result in confidence intervals roughly half the size of IV. We conclude with a discussion of practical guidance for improving precision through the design and analysis of experiments when noncompliance is a concern, including how researchers can determine if the exclusion restriction or PI is appropriate for their experiment.

2. Approaches to addressing noncompliance

We begin by formalizing the issue of noncompliance using principal stratification, a framework for dealing with variables that lie causally between treatment assignment and the outcome, such as noncompliance, attrition, or causal mediators (Frangakis and Rubin, 2002). Assume a finite-sample of $N = 2n$ units, with n units assigned to each of encouragement and control.[2] We assume $Z_i \in \{0, 1\}$ is a completely randomized encouragement. Let $Y_i(z)$ be the potential outcome for unit

[1]A search for "principal score" and "principal ignorability" in the *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, and *Political Analysis*, through Google Scholar, turned up no results. A search for "compliance score" returns one relevant citation, Aronow and Carnegie (2013), which focuses on a different estimand.

[2]See Appendix A-2.1 for a discussion of blocks of arbitrary size and an imbalanced number of units assigned to treatment and control.

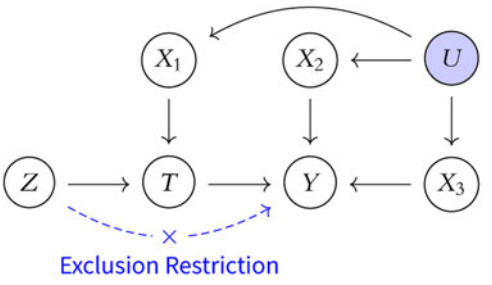

Figure 1. Data Generating Process for Simulations. The dashed line between Z and Y represents a hypothetical path that would occur under a violation of the exclusion restriction. When this path does not exist, the exclusion restriction holds.

i when assigned to encouragement ($Z_i = 1$) or control ($Z_i = 0$). We invoke the stable unit treatment value assumption (Rubin, 1980)—there exists only one version of treatment that is assigned and received with no interference.

Randomization identifies the intent-to-treat effect (ITT), $\tau_{ITT} = \mathbb{E}[Y_i(1) - Y_i(0)]$, where the expectation is taken over the randomization distribution. This is estimated as the difference-in-means between those assigned to encouragement and control. With perfect compliance, the ITT is equivalent to the average treatment effect, but may not be so under noncompliance (see Hirano *et al.*, 2000). Naively evaluating the effect using an as-treated analysis risks bias due to confounding by compliance status. We discuss an example data generating process that demonstrates this in Figure 1, which forms the basis of our simulations. Researchers interested in evaluating the effect of treatment, rather than encouragement, must rely on additional assumptions.

Let T_i represent whether treatment is received by an individual. $T_i(Z_i)$ is a potential outcome based on encouragement, and as such it is fixed, but unobserved, before treatment. The combination of $T_i(Z_i)$ define our principal strata. We focus on one-way noncompliance and assume that every unit assigned to the control condition complies, formalized as:

Assumption 1 **(Strong Monotonicity)** $T_i(Z_i = 0) = 0 \; \forall \; i$

Assumption 1 rules out the existence of two "principal strata":[3] units where $T_i(1) = 0$ and $T_i(0) = 1$ ("defiers" who take the opposite of their encouragement status) and units for which $T_i(1) = 1$ and $T_i(0) = 1$ ("always-takers" who always receive treatment) (Imbens and Rubin, 2015). In encouragement experiments when control units are not given access to treatment, and thus cannot opt-in to treatment, this assumption holds by design, as is often the case with the canonical GOTV outreach experiments. However, in some trials this may not be guaranteed by design, such as for a unit that will attend a rally regardless of whether or not they are encouraged. Under strong monotonicity with one-way noncompliance, there exist two remaining principal strata defined by the latent compliance status, C_i:

$$C_i = \begin{cases} 1 \text{ (Complier)} & \text{if } T_i(1) = 1 \text{ and } T_i(0) = 0, \\ 0 \text{ (Never} - \text{Taker)} & \text{if } T_i(1) = 0 \text{ and } T_i(0) = 0. \end{cases}$$

The strength of principal stratification comes from assuming C_i is fixed pretreatment; randomization ensures ignorability holds within subgroups defined by compliance status as it would for any pretreatment covariate, such as age or gender. Principal causal effects are defined within principal strata (Ding and Lu, 2017; Feller *et al.*, 2017), including the CACE, which gives the effect of actually receiving treatment among compliers:

$$\text{CACE} = \mathbb{E}[Y_i(1) - Y_i(0) \mid C_i = 1]$$

[3]This is sometimes weakened to $T_i(1) \geq T_i(0) \; \forall \; i$, which only rules out defiers (Angrist *et al.*, 1996).

The CACE is a local effect for compliers, which is policy relevant for treatments that can be targeted to individuals likely to take-up the treatment in the real world, and thus it is important for cost-benefit analyses. It is also important for testing theory, where theorized mechanisms only operate on those who take up treatment in the real world. The CACE could differ substantially from the ATE, and researchers should carefully defend the relevance of this estimand for their broader research goals. In particular, because compliers are likely not representative of the full experimental sample, it is important to understand how these individuals differ from the larger experimental sample (Sovey and Green, 2011; Marbach and Hangartner, 2020), and whether the CACE should be adjusted to better reflect the overall average treatment effect (Aronow and Carnegie, 2013).

With noncompliance, we emphasize that randomization alone no longer identifies the CACE. There exist two alternative substantive assumptions researchers can invoke: the excludability assumption invoked with IV, and the principal ignorability assumption invoked with PSW. Neither assumption is guaranteed by randomization but researchers must rely on at least one for identification of the CACE.

2.1 Exclusion restriction

The most common way that experimentalists address noncompliance is with instrumental variables (IV), which invokes the exclusion restriction for identification. We direct readers to Sovey and Green (2011) for details and only review important aspects below. The exclusion restriction states that the only pathway by which encouragement affects the outcome is through receipt of treatment:

Assumption 2 **(Exclusion Restriction)** $Y_i(Z_i, T_i) = Y_i(T_i)$

Estimation is done using two-stage least squares with robust standard errors (Angrist *et al.*, 1996).

While often invoked, researchers must defend the validity of excludability in their application, as it is not guaranteed by randomization. For example, in Sovey and Green (2011, pg. 199), the authors discuss a design in which units are encouraged to watch a Fox News TV Special, and challenge the researcher to ask, "Could it be that opinion change is induced when a person is invited to watch the TV special, regardless of whether he or she in fact watches?" If the invitation to participate can have an impact on the outcome, such as through priming, or changing the behavior of enumerators implementing treatment, then the exclusion restriction is violated. It is incumbent on researchers to justify their invocation of the exclusion restriction.

While IV provides a consistent estimator of the CACE under the additional assumptions outlined above, when compliance rates are low, estimation may be unstable due to the "weak instrument," making inference difficult. This is the primary concern we wish to address by proposing PI approaches. The variance of the IV estimator is inversely proportional to the probability of compliance meaning that ceteris paribus, as the compliance rate drops, the sample size must increase exponentially to match the efficiency of full compliance (Nickerson, 2005). For example, if the compliance rate is 10%, the sample size must increase 100× for the same efficiency as full compliance.

When the probability of compliance is low, even small violations of the exclusion restriction can lead to significant bias (Gerber and Green, 2012). Furthermore, bias from failed randomizations (Imai, 2005) or differential attrition and non-response, such as may occur with survey outcomes measured after a field experiment (Nickerson, 2005; Montgomery *et al.*, 2018), can be exacerbated by low compliance rates when estimating the CACE.

2.2 Principal ignorability

Recent papers in social statistics have proposed an alternative approach to the exclusion restriction assumption for identifying principal causal effects using a "principal ignorability" (PI)

assumption. We emphasize that this is an alternative, but not interchangeable, approach to identifying the CACE. Rather than an exclusion restriction, PI relies on a conditional ignorability assumption where conditional on observable covariates, principal strata membership is as-if randomized.

Assumption 3 **(Weak Principal Ignorability)**

$$\mathbb{E}(Y_i(0) \mid \mathbf{X}_i, T_i(1) = 1) = \mathbb{E}(Y_i(0) \mid \mathbf{X}_i, T_i(1) = 0) = \mathbb{E}(Y_i(0) \mid \mathbf{X}_i)$$

In short, weak PI states that for units assigned to control, conditional on observable covariates, compliance status when assigned to treatment, and therefore principal strata membership, is unrelated to the outcome. We focus on the "weak" PI assumption in this paper, which identifies the CACE under one-way noncompliance. Ding and Lu (2017) and Feller *et al.* (2017) provide thorough introductions to PI, including extensions to two-way noncompliance.

While PI is not verifiable, compliance status is observable in the group randomly assigned to encouragement; researchers should defend weak principal ignorability by identifying factors related to the compliance mechanism among the group assigned to encouragement. Conditioning on variables highly predictive of compliance in the randomized encouragement group lends credibility to weak PI, since there is less residual variance for unobservables, and randomization ensures these are representative compliers (Stuart and Jo, 2015). Researchers should also evaluate covariate balance between the treated and weighted control group (Ding and Lu, 2017), after using adjustment methods described below. There will always be concerns about remaining unobservable confounders, which researchers should probe the plausible impacts of using sensitivity analyses like those we derive in Appendix A-2.2 and those discussed in Ding and Lu (2017).

To clarify the difference in the exclusion restriction and PI assumptions, imagine a researcher wishes to study the impact of political rally attendance on political participation (adapted from McClendon (2014)). Experimental units are randomly assigned to receive a motivational flyer encouraging rally attendance. The outcome is whether an individual votes in the subsequent election, as reported by the Secretary-of-State. In this setting, principal strata membership is defined by rally attendance. Thus, the exclusion restriction requires the message in the flyer only impacts turnout through rally attendance, and not through any other mechanisms, such as providing information about the election. Weak PI states when units do not receive a flyer, attendance if encouraged is unrelated to their voter turnout given a set of observable covariates such as age, political party, etc. Both assumptions are strong. Perhaps the flyer itself motivates individuals to vote, violating excludability. Weak PI may be violated if, conditional on observable covariates, compliance status if encouraged is still related to turnout under control.

2.2.1 Estimation under principal ignorability

While Assumption 3 is nonparametrically justified, in practice researchers typically estimate the CACE under PI using a "principal score", a balancing score that captures the part of $\mathbf{X}$ related to compliance and the outcome. A principal score is defined as $e_c(\mathbf{x}) = Pr(C_i = 1 \mid \mathbf{X}_i = \mathbf{x})$; it is a balancing score in that, similar to a propensity score, $C_i \perp\!\!\!\perp \mathbf{X}_i \mid e_c(\mathbf{x})$. If weak PI holds conditional on $\mathbf{X}$, it holds conditional on $e_c(\mathbf{x})$ (Feller *et al.*, 2017). Principal score models are typically fit in the encouragement group using logistic regression, mixture models, and other flexible regression approaches; estimated scores, $\widehat{e_c}(\mathbf{x}_i)$, for the control group are constructed using fitted values (for some examples see Mattei and Mealli, 2007; Feller *et al.*, 2017; Lee *et al.*, 2010). Methods developed to improve the estimation of propensity scores can be similarly applied to principal scores. Adjustment can also be done by matching on the propensity score (Jo and Stuart, 2009).

Ding and Lu (2017) find that PSW allows for additional robustness over methods that rely on outcome modeling by bypassing potential misspecification errors in the outcome model.

The principal score weighted estimator for the CACE is:

$$\widehat{\tau}_{PSW} = \frac{1}{n_C} \sum_{i \in \{n_1\}} Y_i \cdot C_i - \frac{\sum_{i \in \{n_0\}} Y_i \cdot \widehat{e}_c(\mathbf{x}_i)}{\sum_{i \in \{n_0\}} \widehat{e}_c(\mathbf{x}_i)},$$

where $\{n_1\}$ and $\{n_0\}$ denote the indices corresponding to the units assigned to encouragement and control, respectively; n_C represents the total number of units in the encouragement group who complied (i.e., $n_C = \sum_{i \in \{n_1\}} C_i$). We discuss the variance estimator in Appendix A-5.

Alternative approaches to estimation within the broader principal stratification literature include bounding the unobservable quantity $\mathbb{E}(Y_i(0) \mid C_i = 1)$ (see Zhang and Rubin, 2003; Lee, 2009; Miratrix *et al.*, 2018; Knox *et al.*, 2020; Duarte *et al.*, 2023, for some examples), imposing further assumptions of the conditional independence structure between covariates and outcomes within principal strata (Ding *et al.*, 2011; Mealli *et al.*, 2016), or using more complex model-based estimation strategies (i.e., Esterling *et al.*, 2011; Mattei *et al.*, 2013). However, these approaches either do not address point estimation, or rely on stronger identifying or modeling assumptions, which can lead to model misspecification issues (Feller *et al.*, 2016). As such, we will focus our attention primarily on the PSW approach, which, given estimated principal scores, allows for the nonparametric estimation of the CACE.

3. Improving precision through design

It is well known that blocking can lead to efficiency gains (see Horiuchi *et al.*, 2007; Imai *et al.*, 2008, for a discussion). This literature emphasizes the need to stratify on blocking variables that explain variation in the outcome, and that blocking will generally improve precision (Pashley and Miratrix, 2021a). We show that blocking on variables related to compliance, in addition to the outcome, can be used to more efficiently estimate the CACE. Regardless of whether the exclusion restriction or PI is invoked for identification, blocking allows for more stable estimates and should be considered whenever possible.

In Appendix TA-1 we formalize the precision gains from blocking for the principal ignorability estimators and IV. Additionally, we show that when blocking is possible and PI holds, our ex-ante justified blocked difference-in-means estimator, discussed below in Section 3.1, is more precise than the PSW and IV estimators, thereby demonstrating the power of design-based considerations over post-hoc adjustments. In our simulations and application, we demonstrate significant gains to the PI approaches over IV with complete randomization, and modest gains over blocked-IV.

Extending the findings of Miratrix *et al.* (2013), we show that blocking can result in significant efficiency gains in the IV setting. Previous literature has highlighted the deterioration of the IV estimator's performance under a weak instrument (i.e., high rates of non-compliance), specifically with respect to inflated standard errors and finite sample bias (Bound *et al.*, 1995). Blocking can help offset the instability from a weak instrument. Therefore, even if researchers are utilizing the exclusion restriction as their identifying assumption, accounting for compliance (and/or outcome variation) during the design stage can help offset the precision loss associated with high rates of non-compliance.

When compliance status can be measured among both treated and control units, researchers can also consider a placebo-controlled design in which compliance is measured pretreatment and used as an inclusion criterion for the experiment (Nickerson, 2005; Broockman *et al.*, 2017). We provide more detail in Appendix A-1. As outlined in the literature, these designs can provide significant gains to precision over standard IV, which we see in our simulations, although the gains

are not as significant as the PI approaches (see Table A-3.2). Placebo-controlled designs also benefit from blocking, however because blocking does not guarantee balance across the complier units, the relative precision gains from blocking for the placebo-controlled estimator are less notable than for the other estimators, as shown in the Table A-3.2.

3.1 A special case: principal ignorability at the design stage

We now discuss how to combine the PI assumption with a block-randomized design. To begin, we introduce a modified version of PI, which we refer to as block principal ignorability (Block-PI).

Assumption 4 **(Block Principal Ignorability)** $Y_i(0) \perp\!\!\!\perp C_i \mid B_i$

The implication of Block-PI[4] is that the researcher should construct blocks based on covariates $\mathbf{X}_i$ (or in high dimensional settings, a principal score $e_c(\mathbf{x}_i)$) related to the outcome and compliance. Assuming a matched-pair design (i.e., blocks of size 2), a natural way to estimate the CACE is to limit the analysis to pairs in which the unit receiving encouragement complies[5] and compute the difference-in-means over these pairs. We refer to this as the block-DIM estimator, formalized as:

$$\widehat{\tau}_B = \frac{1}{\sum_{b=1}^{B} C_b} \sum_{b=1}^{B} C_b \widehat{\tau}_b, \tag{1}$$

where $\widehat{\tau}_b$ is the difference-in-means estimated within the bth block, C_b is an indicator that takes on a value of 1 if the bth block contains a complier in the encouragement group, and B is the total number of blocks. We show our estimator is consistent, extend it to blocks of arbitrary size, and derive a conservative variance estimator in Appendices TA-2.1 and A-5.

Conceptually, Block-PI is very similar to the standard weak PI assumption; since the blocks are constructed using pretreatment covariates $\mathbf{X}_i$, we are, in effect, creating blocks within which we assume weak PI holds. However, there are two advantages to the proposed design-stage approach over principal score weighting: it relies on a weaker set of modeling assumptions, and it ensures balance within the sample. First, blocking provides a nonparametric alternative to PSW by directly controlling for the covariates $\mathbf{X}_i$, without invoking additional, untestable assumptions about the underlying principal score model.[6] Second, PSW ensures balance on $\mathbf{X}$ in expectation, but not within individual samples. Block-randomization reduces finite-sample variation because it guarantees balance within blocks for every realization of treatment. Thus, within each randomization, B_i will be orthogonal to Z_i, eliminating the need for adjustment on variables exactly balanced by blocking. Under regular PI, we can only assume that $\mathbf{X}_i$ is independent of treatment assignment in expectation. In Appendix TA-1.2.1 we formally derive when the block-DiM will have lower variance than PSW. Intuitively, the reduction comes from a term that captures the average variance of the compliance rate within a given block. If we can closely "match" these units with similar probability of compliance within blocks, we obtain more precision gains. Additionally, in Appendix A-2.2 we derive the bias due to violations of Block-PI, and introduce a two parameter sensitivity analysis to evaluate robustness of the CACE estimate.

We note that this approach is similar in spirit to that of Jo and Stuart (2009), which does ex-post pair-matching using an estimated propensity score, particularly if blocks are created

[4]Identification of the CACE only requires mean exchangeability, similar to weak PI, but we require full independence for the variance results in the technical appendix.

[5]This is conceptually similar to the placebo-controlled design. The included encouraged units will be the same but the controls come from the corresponding blocks, and are representative of compliers under block PI.

[6]When blocking on a principal score, the parametric assumptions underlying the principal scores still matter.

using the principal score. As with PSW, this method relies on PI, but justifies it using a balancing score based on the propensity to comply; this imposes parametric modeling assumptions.[7] The advantage of block PI is that it incorporates the standard PI assumption *into the design stage*, emphasizing the need to block not just on variables related to the outcome, as one would do to improve precision, but also on variables that render compliance ignorable, as required for identification.

3.2 Variable selection for blocking designs

An important question is how to select what variables to block on. While *identification* under PI indicates researchers should block on covariates that are related to compliance, *precision* is largely dependent on blocking on covariates related to variation in the outcome, consistent with the existing literature on when to expect precision gains from blocking. We suggest researchers should block on covariates related to compliance *and* the outcome. When a researcher has existing data from a similar context, she can do statistical variable selection for determining the important predictors of compliance and the outcome. We discuss one approach, using permutation-based variable importance plots for Random Forest models, in our application in Section 5. An in-depth review of statistical methods for variable selection is beyond the scope of this manuscript.

When existing data is unavailable for researchers to investigate predictors of compliance and outcome, theory must drive variable selection. For example, in a GOTV experiment, if number of children in a household is predictive of a voter's likelihood to answer a GOTV canvassing attempt, researchers should include it as a blocking variable to meet PI. This is true even if it is not predictive of turnout, the primary outcome, especially after accounting for age and previous vote history, which researchers might already block on given their strength in predicting turnout. This demonstrates the difference in what variables researchers should consider for identification, i.e., those related to compliance, versus those they should consider for precision, i.e., those related to the outcome. In Appendix A-2.2 we discuss sensitivity analyses that can help researchers evaluate the robustness of the blocking sets under block principal ignorability.

Finally, it is worth emphasizing that when pre-existing data is unavailable for blocking, researchers will have to rely on ex-post methods at the analysis stage to estimate the CACE. In our simulations, PSW is more efficient than IV for estimating the CACE when compliance is low. Alternatively, poststratification can be used in the analysis stage to recover many of the benefits of blocking for many estimators, including IV (Miratrix *et al.*, 2013; Pashley *et al.*, 2023). However, researchers must *measure* variables related to compliance in order to estimate principal scores or perform poststratification. When blocking before random assignment is not an option, researchers should still use theory to determine what covariates meet principal ignorability, and they should measure these with, or append them to, their experimental data whenever possible.

4. Simulation studies

In this section, we evaluate the performance of the IV, PSW, and blocked difference-in-means estimators for the CACE using simulations. Figure 1 provides a graphical depiction of the data-generating process, and details are provided in Appendix A-3.

To evaluate the impact of the block PI assumption, and violations thereof, we consider the following scenarios: (1) blocking on all compliance-related variables (X_1, "*compliance*"), (2) blocking on all compliance-related and a subset of outcome-related variables (X_1 and X_2, "*compliance + outcome*"), (3) blocking on a subset of outcome-related variables (X_2, "*outcome*"). While both X_2 and X_3 affect the outcome, we do not block on X_3, such that the blocks do not perfectly

[7]When the number of strata grows, block PI may impose additional parametric assumptions through decisions about how to coarsen variables.

explain outcome variation. This represents the real-world setting, in which researchers have not accounted for all variation in the data generating process. We run 5000 simulations with sample size $n = \{1000,\ 2000,\ 5000,\ 10000\}$.

Within each blocking scenario, we consider the following five estimators: under the exclusion restriction, we consider the IV estimator with complete randomization (IV) and blocked IV estimator (IV Block); under PI we consider the ex-ante justified blocked difference-in-means (Block DiM) and the ex-post justified principal score weighted estimator under complete randomization (PSW) and blocked principal score weighted estimator (PSW Block). We include results for the as-treated difference-in-means (AT) for comparison, and provide results for the placebo-control estimator (Placebo) and the blocked placebo-control estimator (Placebo Block) in Appendix Table A-3.2. Pair blocking was performed using the quickblock package in R (Higgins *et al.*, 2016). IV estimators were estimated using the estimatr package (Blair *et al.*, 2022). Principal scores were estimated using a logistic regression, where the covariates in the regression are the same as the covariates used in blocking. As shown in Figure 1, the exclusion restriction holds under all scenarios, and PI holds when blocking and controlling for compliance-related variables (i.e., Scenarios 1 and 2).

Table 1 presents bias and MSE results for all of the simulation scenarios. Overall, we see large improvements in precision for the IV and principal ignorability estimators from using a blocking design, although the efficiency gains the PI approaches is more notable under our low compliance regime, even when PI does not hold.

We first consider the scenarios under which PI holds, Scenario 1 ("*compliance*") and Scenario 2 ("*compliance + outcome*"). As expected, the blocked difference-in-means, which incorporates PI into the ex-ante design, and the ex-post PSW estimators are unbiased. We see significant precision gains from blocking on variables related to the outcome in Scenario 2 ("*compliance + outcome*"), where the MSE for the PI estimators is between 10% and 30% of the MSE from blocking on compliance variables alone. The gains for the blocked IV estimator are even more significant, with the MSE of blocking on compliance and outcome related variables is 2% that of blocking on compliance alone.

Table 1. Summary of simulation results

Simulation Results												
	MSE						Bias					
Sample Size	As Treated	IV	IV (Block)	Block DiM	PSW	PSW (Block)	As Treated	IV	IV (Block)	Block DIM	PSW	PSW (Block)
Scenario 1: Block on Compliance-Related Variables												
1,000	21.68	94.33	67.99	8.62	8.08	6.17	3.53	−0.13	0.10	0.00	−0.02	−0.01
2,000	16.75	44.37	33.68	4.20	3.81	3.08	3.51	−0.10	0.04	−0.00	−0.02	0.00
5,000	14.25	18.25	13.39	1.69	1.52	1.20	3.54	0.03	0.04	−0.02	0.00	−0.02
10,000	13.39	8.60	6.73	0.86	0.74	0.61	3.54	0.03	0.01	0.01	0.01	−0.00
Scenario 2: Block on Compliance and Outcome-Related Variables												
1,000	21.83	95.01	1.60	0.88	3.23	0.96	3.56	−0.02	0.02	−0.00	−0.00	−0.02
2,000	17.26	45.38	0.71	0.42	1.52	0.47	3.60	0.13	0.03	0.01	0.04	−0.00
5,000	14.33	16.88	0.25	0.16	0.58	0.18	3.55	0.12	−0.01	−0.01	0.02	−0.01
10,000	13.46	8.66	0.12	0.08	0.29	0.09	3.55	−0.02	0.00	−0.00	0.01	−0.01
Scenario 3: Block on Outcome-Related Variables												
1,000	21.01	93.42	1.27	0.83	2.79	0.95	3.53	−0.13	−0.00	0.15	0.17	0.14
2,000	16.73	45.74	0.63	0.44	1.40	0.49	3.52	0.07	0.01	0.16	0.17	0.16
5,000	14.17	17.86	0.25	0.19	0.57	0.21	3.53	−0.03	−0.00	0.16	0.15	0.15
10,000	13.26	8.92	0.12	0.11	0.31	0.12	3.52	0.02	0.00	0.16	0.16	0.16

The compliance rate is set at 10%. The simulation is run across varying sample sizes and varying blocking variables, for 5000 total iterations for each sample size and blocking scenario. The default is a complete randomization design, with blocked designs denoted as such. We highlight the lowest MSE estimator in each scenario.

We now consider scenarios where PI does not hold. In Scenario 3 ("*outcome*"), we block on a variable that is related to the outcome and is correlated with compliance at $\rho = 0.5$. This reduces bias by about 95%, relative to the as-treated estimator, but, as expected, does not eliminate bias in the Block DiM and PSW estimators. Because the exclusion restriction still holds, the IV estimator is unbiased, and the blocked IV estimator shows significant improvements in precision over the IV estimator under complete randomization. This emphasizes the flexibility and advantage of using a blocked design, even for estimators that do not rely on principal ignorability. Despite the fact that IV is unbiased, the PI estimators still have lower MSE in our low compliance regime, demonstrating the bias-variance tradeoff researchers must make when choosing between IV and the PI estimators, particularly with low compliance rates.

We note that the efficiency gains of the blocked estimators in comparison to the other estimators is most noticeable when the sample size is small. Our simulation has a low compliance rate, where the efficiency of the IV estimator (under complete randomization) deteriorates rapidly. For example, whenever we block on outcome related variables (i.e., in Scenario 2 or 3), both the blocked difference-in-means estimator, which has the lowest MSE, and the blocked IV estimator are more efficient with a sample size of 1000 than the IV estimator under complete randomization with a sample size of 10,000. When blocking on variables related to compliance (i.e., in Scenario 1 and Scenario 2), the researcher can use either PI or the exclusion restriction for identification of the CACE, and see significant efficiency gains. This shows that, regardless of which identifying assumption the researcher relies on, blocking has meaningful advantages for precision.

4.1 When to expect gains from principal ignorability or the exclusion restriction

Regardless of approach, the necessary identifying assumptions, including the exclusion restriction and PI, are inherently unverifiable. As with any ignorability assumption, it is unlikely that PI holds exactly in practice. Similarly, the exclusion restriction may not hold when it is not justified by design. Therefore, we allow for violations of these identifying assumptions in our simulations to evaluate the bias-variance trade offs researchers face.

To start, we consider simulations in which we allow the violation of principal ignorability to range from minor to extreme. To do so, we replace blocking on X_1 and X_2 in our simulations above with proxy variables that range from completely uncorrelated to perfectly correlated with our original variables. Doing so allows us to change the degree of violation of PI by changing the level of correlation with the original variable. More details can be found in Appendix A-3.2.

Figure 2 shows that, even in the face of violations of the identifying assumption, PI approaches can have significantly lower mean-squared error (MSE) than IV. The variance of IV increases exponentially as compliance rates drop, thus even with PI is violated to some degree, the variance reduction can dominate the squared-bias in many practical applications when researchers can adequately explain compliance or the outcome, resulting in improvements to MSE. For example, in our simulations, if the correlation of our proxy variables is above 0.4, then PI approaches dominate on MSE even though they are biased.

Of course, there is always a concern about choosing a more precise estimator in the face of bias. While our simulations show that, in terms of MSE, this bias-variance trade-off is warranted, we strongly suggest that researchers combine all PI analyses with the sensitivity analysis described in Appendix A-2.2 and Ding and Lu (2017). This allows researchers to assess the impact of violations of PI, determine if potential bias is of substantive concern, and evaluate whether the bias-variance trade-off is worthwhile.

In Appendix A-3.1, we show that when PI holds and the exclusion restriction fails, the PI approaches strictly dominate the IV approaches. Generally, when the exclusion restriction is violated, IV estimation incurs a large degree of bias, while already exhibiting larger variance and thus, larger MSE, than the PI approaches in our low compliance regime.

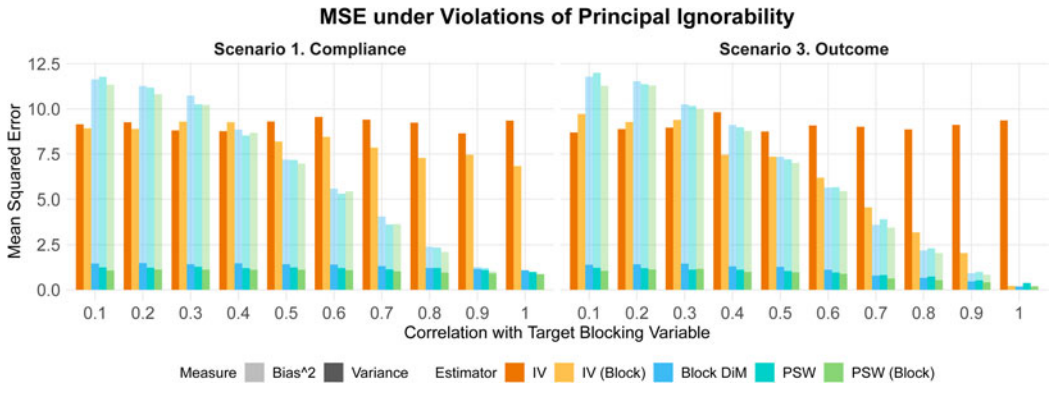

Figure 2. MSE of principal ignorability and exclusion restriction approaches with varying degrees of violation of the principal ignorability assumption. To help identify the drivers of error, we have decomposed the mean squared error into the squared bias and variance components.

Given the gains to precision for the PI are most prominent in low compliance regimes, a natural question is how bad noncompliance must be before the loss in precision for IV estimators is substantial enough to consider PI approaches, which may impose stronger assumptions? To evaluate this, we return to our original simulations set-up, this time varying the compliance rate from 10% to 90%.

Results evaluating the impact of the compliance rate are presented in Figure 3. As evident, when compliance rates are high, such as above 70%, the IV and PSW estimators have similar performance, with PSW showing only slight gains. Similarly, when blocking on variables related to outcome, the blocked IV and PI estimators have similar standard errors with compliance above 30%. Blocking on variables related to just compliance, however, does not provide the same gains to the blocked IV estimator in low compliance regimes as it does for the PI approaches, which show noticeable gains with compliance at or below 50%.

5. Empirical evaluation: Get-Out-the-Vote

To evaluate the advantages of PI for addressing noncompliance, we now turn to a re-evaluation of the Get-Out-the-Vote experiments by Green *et al.* (2003). The original study conducted voter mobilization experiments across six different cities to assess the effect of personal canvassing on verified voter turnout in the November 6, 2001 election. About half of the units within each experimental site were assigned to a personal canvasser contact, with controls receiving no attempt. The overall compliance rate in the encouragement group was 29%, with compliance rates in individual sites ranging from 14% (Columbus) to 45% (Raleigh).

The original experiment did not include block-randomization and used IV to estimate the CACE. In order to mimic a block-randomized design, we construct blocks by performing 1:1 nearest neighbor matching (Sekhon, 2011) using all available covariates;[8] we present a variable selection method for limiting to covariates predictive of the outcome and compliance in Appendix A-4. We are left with 17,442 matched units, where matched pairs constitute the blocks in our analyses.[9]

[8]Since encouragement was randomized our matching algorithm does not impose a selection-on-observables assumption, however identification of the CACE under the block-DiM does. In some sense, the approach is similar to that of Jo and Stuart (2009), where matching is done using the full covariate profile rather than a dimension-reducing propensity score. See Gerber and Green (2012) §4.5 for a more in-depth discussion of the similarity between blocking and analysis-stage covariate adjustment with large samples similar to our analysis.

[9]We conduct our analyses for complete randomization on a comparable sized dataset by randomly dropping about 15% of control units. This makes the results not directly comparable to the original point estimates, but results are substantively similar.

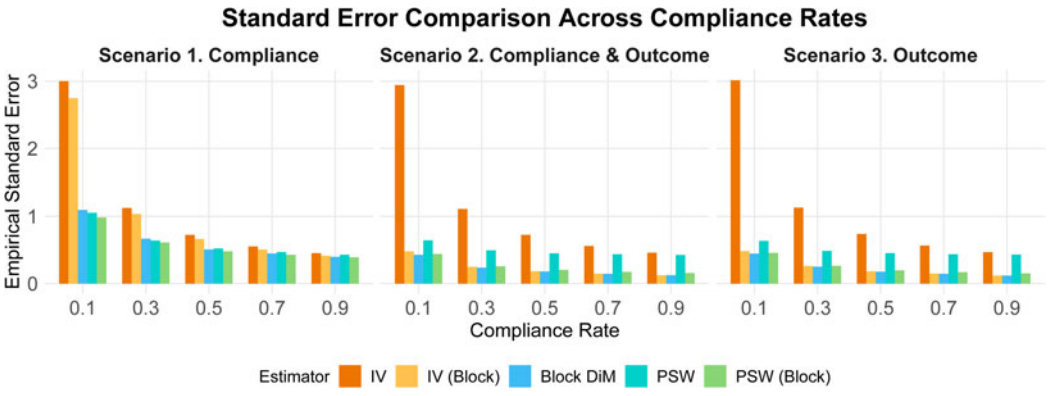

Figure 3. Standard error of estimators using DGP described in Figure 1 with increasing levels of compliance.

We note that we predict compliance in the encouragement group using our covariates, including measures for vote history, age, and family size, with 72–95% accuracy across sites, and that this set includes some of the most important predictors of both compliance (age) and turnout (age and vote history). This lends some credibility to the PI assumption (see Appendix A-4.2) since observables explain compliance well in the treated group. However, researchers should rightfully be concerned there remain unobserved confounders, such as employment status. We also highlight that our sensitivity analysis in Appendix A-2.2 indicates robustness to plausible violations from such confounders, indicating that while there may be residual bias, it is unlikely to overturn our conclusions. We evaluate five estimators, including two justified by the exclusion restriction, (1) the IV estimator under complete randomization, (2) the block-IV estimator, and three justified by PI, (3) the PSW estimator under complete randomization, (4) the block-PSW estimator, and (5) the ex-ante justified block-DIM estimator. The principal scores are estimated using logistic regression including the same variables used to form blocks. Standard errors are estimated using robust standard errors (HC2) under complete randomization and cluster-robust standard errors (CR2) under block-randomization.

We highlight two main takeaways: (1) there are significant gains in precision for the PI estimators over IV, with minimal evidence of residual bias, and (2) blocking improves the precision of all approaches. A visualization of the results is provided in Figure 4 with numerical values in Table A-4.5. Gains to precision are presented in Table 2. We first note that across all of the experimental sites, the PI approaches produce more precise estimates than IV. The PSW estimator, under complete randomization, has a standard error 53% that of the IV estimator under complete randomization. With block-randomization, the block-PSW and block-DIM estimator exhibit even larger gains to precision when compared to the standard IV estimator with complete randomization. In some cases, the significant gains to precision from the PI approaches change the statistical significance of the results, thus showcasing the opportunity for more precise estimates using PI. However, these gains should be considered in light of the fact that in this application, the exclusion restriction is likely to hold and therefore IV should be unbiased. In contrast, whether or not principal ignorability is a valid identifying assumption requires researcher judgment. Recall our sensitivity analysis indicates robustness to plausible violations supporting minimal bias. The similarity in the point estimates across methods provides additional credibility that the violations to PI, if they exist, are likely small. Therefore, we should expect a reduction in MSE from relying on principal ignorability, driven by a reduction in variance, given the low compliance rate of 29%.

Second, we highlight that regardless of which identifying assumption researchers use, blocking on variables related to the outcome and compliance improves precision. More specifically, block-

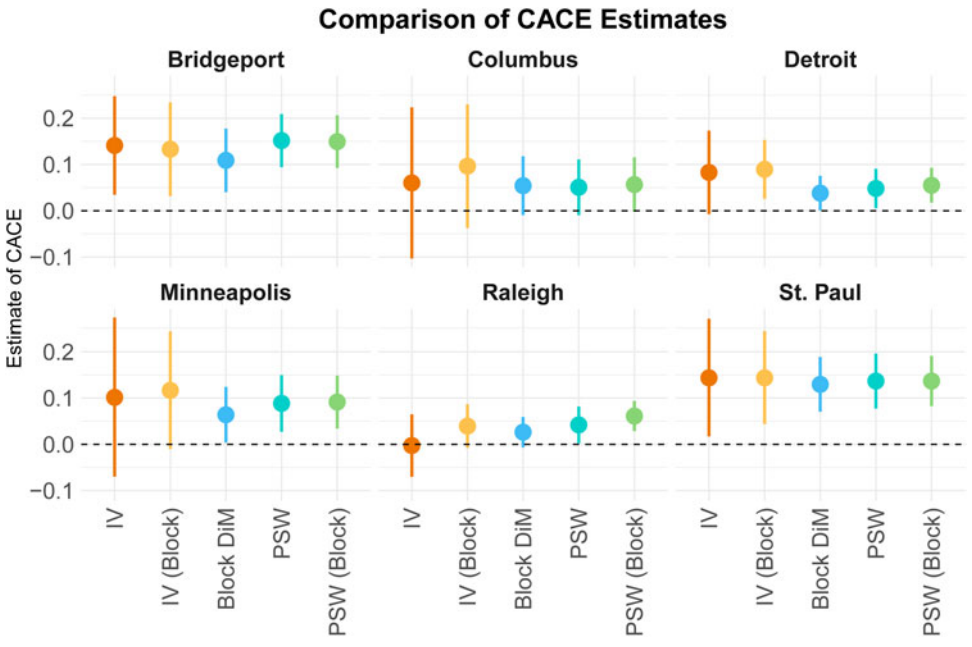

Figure 4. Plot of the point estimates and 95% confidence intervals across experimental sites.

Table 2. Percentage reduction in estimated standard error relative to IV estimator under complete randomization

City	IV (Block)	PSW	PSW (Block)	Blocked DIM
Bridgeport	4.8%	46.0%	46.3%	35.4%
Columbus	18.3%	63.1%	64.0%	61.1%
Detroit	29.2%	52.8%	58.3%	58.5%
Minneapolis	26.1%	64.2%	66.6%	64.9%
Raleigh	29.9%	41.1%	51.6%	51.2%
St. Paul	21.0%	53.0%	57.2%	53.2%

randomization reduces the standard error of the IV estimator considerably. In sites where covariates explain a substantial amount of the variation in the outcome (i.e., Detroit, Raleigh), blocking results in a nearly 30% reduction in the IV standard error. Even in experimental sites where the covariates are less explanatory (i.e., Bridgeport), blocking still results in a 5% reduction. Similarly, block-randomization improves precision for the PSW estimator over complete randomization, although gains are less noticeable than for IV.

6. Practical guidance and discussion

There are many practical considerations for designing and analyzing experiments with noncompliance. This article has introduced a framework, including a set of identifying assumptions and design considerations that aim to improve precision of estimators for the CACE, particularly when noncompliance is significant to severe and IV estimation may exhibit high variance. In particular, we suggest researchers implement a block-randomized design, regardless of estimation strategy. We then provide an alternative identifying assumption, PI, in service of discussing and developing more precise estimators for the CACE. In particular, we combine the ex-ante blocking design with PI into block-principal ignorability, and introduce a simple block-difference-in-means

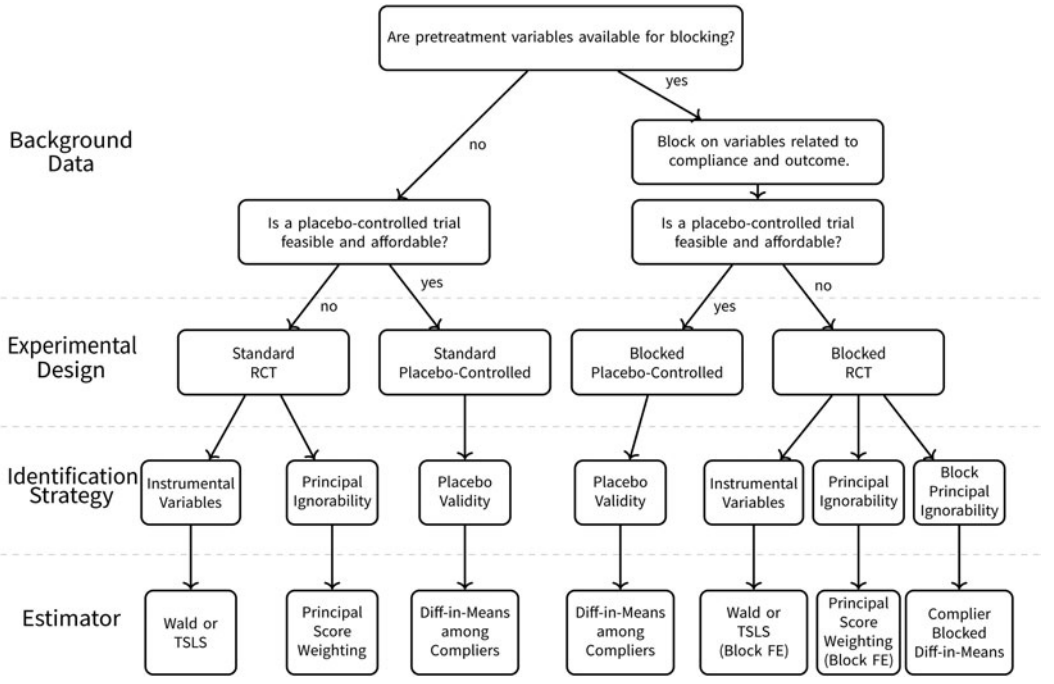

Figure 5. Data, design, and analysis considerations for estimating the complier average causal effect in experiments with non-compliance.

estimator with desirable properties. We summarize our suggestions for designing and analyzing experiments in Figure 5.

Based on the discussion in this article, we first emphasize that researchers should, where possible and regardless of identifying assumptions and estimation strategy, use a block-randomized design, blocking on variables related to both outcome and compliance when the CACE is the quantity of interest. Researchers could also consider a placebo-controlled design which improves precision, particularly when compliance is less than 50% (Gerber and Green, 2012). This can reduce the cost of the experiment for some designs (Broockman *et al.*, 2017). In our simulations, the gains to blocking with placebo-controlled trials are less noticeable than for other approaches.

A natural question is what to include in the blocking design. The PI assumption indicates researchers should block on variables that explain compliance, as identification requires that compliance be rendered ignorable. Standard practice suggests they should consider variables predictive of the outcome of interest for greatest precision gains. We strongly suggest researchers do both, and evaluate the set using the tools described below. When previous data exists on compliance or outcomes, researchers can use theory or statistical variable selection methods to inform the blocking design. When existing data are unavailable, theory must drive variable selection. We emphasize that when block-randomization is not an option, researchers should still use theory to justify PI, and they should *measure* these variables related to compliance and the outcome with, or append them to, their experimental data for use in adjustment.

In addition to the blocking design discussed above, we introduced the principal ignorability assumption. To identify the CACE, researchers must rely on additional substantive assumptions not guaranteed by randomization. Researchers can invoke either the exclusion restriction or the PI assumption, both of which are untestable. Researchers should carefully consider which of these alternative approaches is most appropriate for their experiment. The exclusion restriction is justifiable when control units cannot access treatment but should be approached cautiously

otherwise. Additionally, when compliance rates are high, the IV estimator does not suffer from significant loss to precision and may be most appropriate, as shown in Figure 3.

However, when compliance rates are low, or the exclusion restriction is implausible, researchers may wish to invoke PI—an alternative, but not interchangeable assumption. While there is no statistical criterion that can guarantee PI holds, researchers can bolster the PI assumption by: (1) evaluating the fit of the principal scores (Stuart and Jo, 2015), (2) evaluating balance of pretreatment covariates between compliers and weighted noncompliers (Ding and Lu, 2017; Feller *et al.*, 2017) or balance in coarsened blocks, and (3) conducting sensitivity analyses (Appendix A-2.2 and Ding and Lu, 2017). Additionally, even when violated to some degree, principal ignorability may still be preferable as the blocked-PSW, PSW, and the blocked-DIM estimators may outperform IV in terms of MSE when compliance is very low, as seen in Section 4.1. Where both assumptions are plausible, researchers can use both methods for a robustness check.

We primarily focus on one-way noncompliance, however PI can be extended to two-way noncompliance to address additional compliance types. See Feller *et al.* (2017) for a discussion of the complications with estimation of principal scores under two-way noncompliance, when compliers are no longer directly observable in the control group. Finally, we note that, as with IV, the block-DIM and the PSW estimators can easily incorporate additional covariates for regression adjustment with estimation via regression or weighted regression, respectively.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2023.38. To obtain replication material for this article, https://doi.org/10.7910/DVN/RZHOI1

Acknowledgements. The authors would like to thank Graeme Blair, Thad Dunning, Avi Feller, Chad Hazlett, Reid Oda, Abby Wood, and the members of the UCLA Causal Inference Reading Group for their valuable feedback. Melody Huang is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2146752. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation. Melody Huang was also supported in part by AFOSR MURI grant #FA9550-22-1-0380.

Competing interest. None.

References

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

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

Blair G, Cooper J, Coppock A, Humphreys M and Sonnet L (2022) estimatr: Fast Estimators for Design-Based Inference. Available at https://declaredesign.org/r/estimatr/, https://github.com/DeclareDesign/estimatr.

Bound J, Jaeger D and Baker R (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* **90**, 443–450.

Broockman D, Kalla J and Sekhon J (2017) The design of field experiments with survey outcomes: a framework for selecting more efficient, robust, and ethical designs. *Political Analysis* **25**, 435–464.

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

Ding P, Geng Z, Yan W and 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.

Duarte G, Finkelstein N, Knox D, Mummolo J and Shpitser I (2023) An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association* **just-accepted**, 1–25.

Esterling K, Neblo M and Lazer D (2011) Estimating treatment effects in the presence of noncompliance and nonresponse: the generalized endogenous treatment model. *Political Analysis* **19**, 205–226.

Feller A, Greif E, Miratrix L and Pillai N (2016) Principal stratification in the twilight zone: weakly separated components in finite mixture models. *arXiv preprint arXiv:1602.06595*.

Feller A, Mealli F and Miratrix L (2017) Principal score methods: assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics* **42**, 726–758.

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

Gerber A and Green D (2012) *Field experiments: Design, analysis, and interpretation*. New York: WW Norton.

Green D, Gerber A and Nickerson D (2003) Getting out the vote in local elections: results from six door-to-door canvassing experiments. *The Journal of Politics* **65**, 1083–1096.
Higgins MJ, Sävje F and Sekhon JS (2016) Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences* **113**, 7369–7376.
Hirano K, Imbens G, Rubin D and Zhou X-H (2000) Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69–88.
Horiuchi Y, Imai K and Taniguchi N (2007) Designing and analyzing randomized experiments: Application to a Japanese election survey experiment. *American Journal of Political Science* **51**, 669–687.
Imai K (2005) Do Get-Out-the-Vote calls reduce turnout? The importance of statistical methods for field experiments. *American Political Science Review* **99**, 283–300.
Imai K, King G and Stuart E (2008) Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* **171**, 481–502.
Imbens G and Rubin D (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.
Jo B and Stuart E (2009) On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine* **28**, 2857–2875.
Knox D, Lowe W and Mummolo J (2020) Administrative records mask racially biased policing. *American Political Science Review* **114**, 619–637.
Lee D (2009) Training, wages, and sample selection: estimating sharp bounds on treatment effects. *The Review of Economic Studies* **76**, 1071–1102.
Lee B, Lessler J and Stuart E (2010) Improving propensity score weighting using machine learning. *Statistics in Medicine* **29**, 337–346.
Marbach M and Hangartner D (2020) Profiling compliers and noncompliers for instrumental-variable analysis. *Political Analysis* **28**, 435–444.
Mattei A and Mealli F (2007) Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics* **63**, 437–446.
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.
McClendon GH (2014) Social esteem and participation in contentious politics: A field experiment at an LGBT pride rally. *American Journal of Political Science* **58**, 279–290.
Mealli F, Pacini B and 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, Sekhon J and Yu B (2013) Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 369–396.
Miratrix L, Furey J, Feller A, Grindal T and Page LC (2018) Bounding, an accessible method for estimating principal causal effects, examined and explained. *Journal of Research on Educational Effectiveness* **11**, 133–162.
Montgomery J, Nyhan B and Torres M (2018) How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science* **62**, 760–775.
Nickerson D (2005) Scalable protocols offer efficient design for field experiments. *Political Analysis* **13**, 233–252.
Pashley N and Miratrix L (2021a) Block what you can, except when you shouldn't. *Journal of Educational and Behavioral Statistics* **47**, 69–100.
Pashley N and Miratrix L (2021b) Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics* **46**, 271–296.
Pashley NE, Keele L and Miratrix LW (2023) Improving instrumental variable estimators with post-stratification *arXiv preprint:arXiv:2303.10016*.
Rubin D (1980) Randomization analysis of experimental data: the fisher randomization test comment. *Journal of the American Statistical Association* **75**, 591–593.
Sekhon J (2011) Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software* **42**, 1–52.
Sovey A and Green D (2011) Instrumental variables estimation in political science: a readers' guide. *American Journal of Political Science* **55**, 188–200.
Stuart E and Jo B (2015) Assessing the sensitivity of methods for estimating principal causal effects. *Statistical Methods in Medical Research* **24**, 657–674.
Zhang J and Rubin D (2003) Estimation of causal effects via principal stratification when some outcomes are truncated by 'death'. *Journal of Educational and Behavioral Statistics* **28**, 353–368.

Cite this article: Hartman E, Huang M (2024). Improving precision through design and analysis in experiments with noncompliance. *Political Science Research and Methods* **12**, 557–572. https://doi.org/10.1017/psrm.2023.38