Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Decomposing Treatment Effect Variation

Peng Ding, Avi Feller & Luke Miratrix

To cite this article: Peng Ding, Avi Feller & Luke Miratrix (2019) Decomposing Treatment Effect Variation, Journal of the American Statistical Association, 114:525, 304-317, DOI: 10.1080/01621459.2017.1407322

To link to this article: https://doi.org/10.1080/01621459.2017.1407322

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 525, 304–317, Theory and Methods
https://doi.org/10.1080/01621459.2017.1407322

Taylor & Francis
Taylor & Francis Group

Decomposing Treatment Effect Variation

Peng Ding[a], Avi Feller[a,b], and Luke Miratrix[c]

[a]Department of Statistics, University of California, Berkeley, CA; [b]Goldman School of Public Policy, University of California, Berkeley, CA; [c]Harvard Graduate School of Education, Cambridge, MA

ABSTRACT

Understanding and characterizing treatment effect variation in randomized experiments has become essential for going beyond the "black box" of the average treatment effect. Nonetheless, traditional statistical approaches often ignore or assume away such variation. In the context of randomized experiments, this article proposes a framework for decomposing overall treatment effect variation into a systematic component explained by observed covariates and a remaining idiosyncratic component. Our framework is fully randomization-based, with estimates of treatment effect variation that are entirely justified by the randomization itself. Our framework can also account for noncompliance, which is an important practical complication. We make several contributions. First, we show that randomization-based estimates of systematic variation are very similar in form to estimates from fully interacted linear regression and two-stage least squares. Second, we use these estimators to develop an omnibus test for systematic treatment effect variation, both with and without noncompliance. Third, we propose an R^2-like measure of treatment effect variation explained by covariates and, when applicable, noncompliance. Finally, we assess these methods via simulation studies and apply them to the Head Start Impact Study, a large-scale randomized experiment. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2016
Revised November 2017

KEYWORDS
Heterogeneous treatment effect; Idiosyncratic treatment effect variation; Noncompliance; Randomization inference; Systematic treatment effect variation

1. Introduction

The analysis of randomized experiments has traditionally focused on the average treatment effect, often ignoring or assuming away treatment effect variation (e.g., Neyman 1923; Fisher 1935; Kempthorne 1952; Rosenbaum 2002). Today, understanding and characterizing treatment effect variation in randomized experiments has become essential for going beyond the "black box" of the average treatment effect. This is clear from the increasing number of articles on the topic in statistics and machine learning (Hill 2011; Athey and Imbens 2016; Wager and Athey 2017), biostatistics (Huang, Gilbert, and Janes 2012; Matsouaka, Li, and Cai 2014), education (Raudenbush and Bloom 2015), economics (Heckman, Smith, and Clements 1997; Crump et al. 2008; and Djebbari and Smith 2008), political science (Green and Kern 2012; Imai and Ratkovic 2013), and other areas.

This article proposes a framework for decomposing overall treatment effect variation in a randomized experiment into a *systematic component* that is explained by observed covariates, and an *idiosyncratic component* that is not explained (Heckman, Smith, and Clements 1997; Djebbari and Smith 2008). In doing so, we make several key contributions. First, we take a fully randomization-based perspective (see Rosenbaum 2002; Imbens and Rubin 2015), and propose estimators that are entirely justified by the randomization itself. This is in contrast to much of the literature on randomization-based methods, where treatment effect variation is typically a nuisance (e.g., Rosenbaum 1999, 2007). Similar to Lin (2013), we show that the resulting estimator is very similar in form to linear regression with interactions between the treatment indicator and covariates. Unlike with linear regression, however, the proposed estimator does not require any modeling assumptions on the marginal outcomes.

Second, we extend these methods from intention-to-treat (ITT) analysis to allow for noncompliance, proposing a randomized-based estimator for systematic treatment effect variation for the local average treatment effect (LATE) in the case of noncompliance (Angrist, Imbens, and Rubin 1996). We show that this estimator is nearly identical to the two-stage least-square estimator with interactions between the treatment and covariates. We believe that this is a particularly novel contribution to the recent literature seeking to reconcile the randomization-based tradition in statistics and the linear model-based perspective more common in econometrics (Abadie 2003; Imbens 2014; Imbens and Rubin 2015).

Armed with these estimators, we turn to two practical tools for decomposing treatment effect variation. The first is an omnibus test for the presence of systematic treatment effect variation. While versions of this test have been proposed previously, largely in the context of linear models (Cox 1984; Crump et al. 2008), our proposed test is fully randomization-based and can also account for noncompliance. The second is to develop and bound an R^2-like measure of the fraction of treatment effect variation explained by covariates. This builds on previous versions proposed in the econometrics literature (Heckman, Smith, and Clements 1997; Djebbari and Smith 2008), again

CONTACT Luke Miratrix lmiratrix@stat.harvard.edu Harvard Graduate School of Education, Cambridge, MA 02138.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

extending results to account for noncompliance. This approach is also closely related to the Oaxaca–Blinder decomposition in economics (Oaxaca 1973; Blinder 1973). See Angrist, Pathak, and Walters (2013) for a recent application that also addresses compliance. Finally, we apply these methods to the Head Start Impact Study, a large-scale randomized trial of Head Start, a federally funded preschool program (Puma et al. 2010). We relegate the technical details and some further extensions to the online supplementary material.

2. Framework for Treatment Effect Variation

2.1. Setup and Notation

Assume that we have n units in an experiment. For unit i, let $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{Ki})^\top \in \mathbb{R}^K$ denote the vector of pretreatment covariates, with the constant 1 as its first component. Let T_i denote the treatment indicator with 1 for treatment and 0 for control. We use the potential outcomes framework (Neyman 1923; Rubin 1974) to define causal effects. Under the stable unit treatment value assumption (Rubin 1980) that there is only one version of the treatment and no interference among units, we define $Y_i(1)$ and $Y_i(0)$ as the potential outcomes of unit i under treatment and control, respectively. The observed outcome, $Y_i^{\text{obs}} = T_i Y_i(1) + (1 - T_i) Y_i(0)$, is quite general and includes continuous, binary, and zero-inflated cases. On the difference scale, the individual treatment effect is $\tau_i = Y_i(1) - Y_i(0)$.

Importantly, this is finite population inference in that we condition on the n units at hand—the potential outcomes are fixed and pretreatment. This differs from super population inference in which some variables or residuals are assumed to be independent and identically distributed (iid) draws from some distribution. See, for example, Rosenbaum (2002), Imbens and Rubin (2015), and Li and Ding (2017). Under the potential outcomes framework, $\{Y_i(1), Y_i(0)\}_{i=1}^n$ are all fixed numbers; the randomness of any estimator comes from the assignment mechanism, which is the distribution of possible treatment assignments $\boldsymbol{T} = (T_1, \ldots, T_n)^\top$. Note that $\text{pr}\{(T_1, \ldots, T_n) = (t_1, \ldots, t_n)\} = \binom{n}{n_1}^{-1}$ if $\sum_{i=1}^n t_i = n_1$.

2.2. Randomization Inference for Vector Outcomes

To set up our overall framework, we first generalize Neyman's (1923) classic results to vector outcomes. We consider a completely randomized experiment, with n_1 units assigned to treatment and n_0 units assigned to control; in total we have $\binom{n}{n_1}$ possible randomizations. We are interested in estimating the finite population average treatment effect on a vector outcome $\boldsymbol{V} \in \mathbb{R}^K$:

$$\boldsymbol{\tau}_{\boldsymbol{V}} = \frac{1}{n} \sum_{i=1}^n \{\boldsymbol{V}_i(1) - \boldsymbol{V}_i(0)\},$$

where $\boldsymbol{V}_i(1)$ and $\boldsymbol{V}_i(0)$ are the potential outcomes of $\boldsymbol{V}$ for unit i. For example, $\boldsymbol{V}$ can be Y or $\boldsymbol{X}Y$. The Neyman-type unbiased estimator for $\boldsymbol{\tau}_{\boldsymbol{V}}$ is the difference between the sample mean vectors of the observed outcomes under treatment and control:

$$\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}} = \bar{\boldsymbol{V}}_1^{\text{obs}} - \bar{\boldsymbol{V}}_0^{\text{obs}} = \frac{1}{n_1} \sum_{i=1}^n T_i \boldsymbol{V}_i^{\text{obs}} - \frac{1}{n_0} \sum_{i=1}^n (1 - T_i) \boldsymbol{V}_i^{\text{obs}}$$
$$= \frac{1}{n_1} \sum_{i=1}^n T_i \boldsymbol{V}_i(1) - \frac{1}{n_0} \sum_{i=1}^n (1 - T_i) \boldsymbol{V}_i(0).$$

The behavior of our estimator, and of our estimators for heterogeneity discussed later, revolve around covariances of vector outcomes. For notation, let $\boldsymbol{A} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n\}$ be a collection of n vectors, with $\bar{\boldsymbol{A}} = n^{-1} \sum_{i=1}^n \boldsymbol{A}_i$ the vector mean, and define the covariance operator on $\boldsymbol{A}$ as

$$\mathcal{S}(\boldsymbol{A}) = \frac{1}{n-1} \sum_{i=1}^n (\boldsymbol{A}_i - \bar{\boldsymbol{A}})(\boldsymbol{A}_i - \bar{\boldsymbol{A}})^\top,$$

which gives the covariance matrix of the n vectors in $\boldsymbol{A}$. For example, $\boldsymbol{A}_i$ can be $\boldsymbol{V}_i(1)$, $\boldsymbol{V}_i(0)$, or $\boldsymbol{V}_i(1) - \boldsymbol{V}_i(0)$.

The following theorem, generalizing the results for scalar outcomes from Neyman (1923), demonstrates that $\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}}$ is unbiased and gives its covariance matrix.

Theorem 1. Over all possible randomizations of a completely randomized experiment, $\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}}$ is unbiased for $\boldsymbol{\tau}_{\boldsymbol{V}}$, with $K \times K$ covariance matrix:

$$\text{cov}(\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}}) = \frac{\mathcal{S}\{\boldsymbol{V}(1)\}}{n_1} + \frac{\mathcal{S}\{\boldsymbol{V}(0)\}}{n_0} - \frac{\mathcal{S}\{\boldsymbol{V}(1) - \boldsymbol{V}(0)\}}{n}. \quad (1)$$

The diagonal elements of this matrix are the variances of the estimators of each component of $\boldsymbol{\tau}_{\boldsymbol{V}}$. The covariance matrix of $\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}}$ depends on the various covariances of the potential outcomes under treatment and control. In particular, the last term depends on the correlation between the potential outcomes $\boldsymbol{V}(1)$ and $\boldsymbol{V}(0)$, and therefore cannot be identified from the observed data. When the individual treatment effects are constant for all components of $\boldsymbol{V}$, the last term in the above covariance matrix vanishes, because then $\mathcal{S}\{\boldsymbol{V}(1) - \boldsymbol{V}(0)\} = \boldsymbol{0}_{K \times K}$. Under this assumption, we can unbiasedly estimate the sampling covariance matrix $\text{cov}(\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}})$ by replacing the covariances of the potential outcomes by the sample analogs:

$$\widehat{\text{cov}}(\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}}) = \frac{\widehat{\mathcal{S}}_1(\boldsymbol{V}^{\text{obs}})}{n_1} + \frac{\widehat{\mathcal{S}}_0(\boldsymbol{V}^{\text{obs}})}{n_0},$$

where

$$\widehat{\mathcal{S}}_t(\boldsymbol{V}^{\text{obs}}) = \frac{1}{n_t - 1} \sum_{i=1}^n I_{(T_i = t)} (\boldsymbol{V}_i - \bar{\boldsymbol{V}}_t^{\text{obs}})(\boldsymbol{V}_i - \bar{\boldsymbol{V}}_t^{\text{obs}})^\top \quad (t = 0, 1) \quad (2)$$

are the sample covariance matrices of $\boldsymbol{V}^{\text{obs}}$ in the treatment and control groups. Without the constant treatment effect assumption, the covariance estimator $\widehat{\text{cov}}(\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}})$ is conservative in the sense that the difference between the expectation of the variance estimator and the true variance is a nonnegative definite matrix. In particular, the diagonal terms of the expected estimator will all be larger than the truth. Letting $K = 1$, the covariance matrices become simple variances, which recovers Neyman's original result.

Using the mathematical framework introduced in the Appendix and in Li and Ding (2017), we can easily generalize Theorem 1 to more complicated experimental designs, for example, cluster-randomized trials (Middleton and Aronow 2015) and unbalanced 2^2 split-plot designs (Zhao et al. 2017).

2.3. Decomposing Treatment Effect Variation

We now apply this general framework to treatment effect variation. We decompose the individual treatment effect, τ_i, via

$$\tau_i = Y_i(1) - Y_i(0) = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad (i = 1, \ldots, n) \tag{3}$$

with $\boldsymbol{\beta}$ being the finite population linear regression coefficient of τ_i on $\boldsymbol{X}_i$, defined by

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^K} \sum_{i=1}^n \left(\tau_i - \boldsymbol{X}_i^\top \boldsymbol{b}\right)^2. \tag{4}$$

Following Heckman, Smith, and Clements (1997) and Djebbari and Smith (2008), we call $\delta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$ the *systematic treatment effect variation* explained by the observed covariates, $\boldsymbol{X}_i$, and call $\varepsilon_i \equiv \tau_i - \delta_i = \tau_i - \boldsymbol{X}^\top \boldsymbol{\beta}$ the *idiosyncratic treatment effect variation* not explained by $\boldsymbol{X}_i$.

More generally, we can view this decomposition in a regression-style framework. Define

$$\boldsymbol{S}_{xx} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top \in \mathbb{R}^{K\times K}, \quad \boldsymbol{S}_{x\varepsilon} = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \boldsymbol{X}_i \in \mathbb{R}^K,$$
$$\boldsymbol{S}_{x\tau} = \frac{1}{n}\sum_{i=1}^n \tau_i \boldsymbol{X}_i \in \mathbb{R}^K,$$

where $\boldsymbol{S}_{xx}$ is nondegenerate, analogous to the usual full-rank assumption in linear models. Also define

$$\boldsymbol{S}_{xt} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i Y_i(t) \in \mathbb{R}^K, \quad (t = 0, 1).$$

These are all finite population quantities, as in they are fixed pre-randomization values. The definition of $\boldsymbol{\beta}$ gives $\boldsymbol{S}_{x\varepsilon} = 0$, that is, ε_i and $\boldsymbol{X}_i$ have finite population covariance zero. Therefore, in the spirit of the agnostic regression framework (e.g., Lin 2013), the systematic component, $\delta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$, is a projection of τ_i onto the linear space spanned by $\boldsymbol{X}_i$, and the idiosyncratic treatment effect, ε_i, is the corresponding residual. The linear projection applies to general outcomes, including the binary case.

Because of our finite population focus, if we observed all the potential outcomes we could immediately calculate all individual treatment effects and apply standard linear regression theory to (3) and obtain $\boldsymbol{\beta}$. In particular, the solution of (4), that is, the ordinary least-square (OLS) solution from regressing τ on $\boldsymbol{X}$, is

$$\boldsymbol{\beta} = \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x\tau} = \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x1} - \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x0} \equiv \boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0, \tag{5}$$

where $\boldsymbol{\gamma}_1 = \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x1}$ and $\boldsymbol{\gamma}_0 = \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x0}$ are the corresponding finite population regression coefficients of the potential outcomes on the covariates. Let $e_i(1) = Y_i(1) - \boldsymbol{X}_i^\top \boldsymbol{\gamma}_1$ and $e_i(0) = Y_i(0) - \boldsymbol{X}_i^\top \boldsymbol{\gamma}_0$ be the residual potential outcomes from the regression of $Y_i(t)$ onto $\boldsymbol{X}$. Our idiosyncratic treatment variation is then the difference of residuals: $\varepsilon_i = e_i(1) - e_i(0)$. In practice, we do not fully observe these components, but we can obtain unbiased or consistent estimates for them as we discuss below.

3. Systematic Treatment Effect Variation for the ITT

3.1. Randomization-Based Estimator

We now turn to estimating $\boldsymbol{\beta}$. As shown in (5), $\boldsymbol{\beta}$ has three components. The first term, $\boldsymbol{S}_{xx}$, is fully observed as all the covariates are observed. Our estimation then depends on the sample analogs of $\boldsymbol{S}_{x1}$ and $\boldsymbol{S}_{x0}$:

$$\widehat{\boldsymbol{S}}_{x1} = \frac{1}{n_1}\sum_{i=1}^n T_i Y_i^{\text{obs}} \boldsymbol{X}_i \in \mathbb{R}^K,$$
$$\widehat{\boldsymbol{S}}_{x0} = \frac{1}{n_0}\sum_{i=1}^n (1 - T_i) Y_i^{\text{obs}} \boldsymbol{X}_i \in \mathbb{R}^K.$$

The $\widehat{\boldsymbol{S}}_{xt}$'s capture how the observed potential outcomes correlate with the covariates. Plug these into (5) to obtain an overall estimate of $\boldsymbol{\beta}$. The randomization of $\boldsymbol{T}$ then justifies the following theorem.

Theorem 2. Under decomposition (3), $\boldsymbol{S}_{xx}^{-1}\widehat{\boldsymbol{S}}_{x1}$ and $\boldsymbol{S}_{xx}^{-1}\widehat{\boldsymbol{S}}_{x0}$ are unbiased estimates of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_0$, respectively. Therefore,

$$\widehat{\boldsymbol{\beta}}_{\text{RI}} = \boldsymbol{S}_{xx}^{-1}\widehat{\boldsymbol{S}}_{x1} - \boldsymbol{S}_{xx}^{-1}\widehat{\boldsymbol{S}}_{x0},$$

is an unbiased estimator for $\boldsymbol{\beta}$ with covariance matrix

$$\text{cov}(\widehat{\boldsymbol{\beta}}_{\text{RI}}) = \boldsymbol{S}_{xx}^{-1}\left[\frac{\mathcal{S}\{Y(1)\boldsymbol{X}\}}{n_1} + \frac{\mathcal{S}\{Y(0)\boldsymbol{X}\}}{n_0} - \frac{\mathcal{S}(\tau\boldsymbol{X})}{n}\right]\boldsymbol{S}_{xx}^{-1}. \tag{6}$$

Here, for example, $\mathcal{S}\{Y(0)\boldsymbol{X}\}$ denotes the covariance operator on new unit-level variables $Y_i(0)\boldsymbol{X}_i \in \mathbb{R}^K$, made by scaling the $\boldsymbol{X}_i$ vector of each unit by $Y_i(0)$, similarly for $\mathcal{S}\{Y(1)\boldsymbol{X}\}$ and $\mathcal{S}(\tau\boldsymbol{X})$. This slight abuse of notation gives formulas less cluttered by subscripts and excessive annotation. As with the vector version of Neyman's formula, the square root of the diagonal of $\text{cov}(\widehat{\boldsymbol{\beta}}_{\text{RI}})$ gives the standard errors of $\widehat{\boldsymbol{\beta}}_{\text{RI}}$.

The covariance formula (6) generalizes the result of Neyman (1923) for the average treatment effect, reducing to Neyman's formula if $\boldsymbol{X}_i = 1$ for all units. We can obtain a "conservative" estimate of $\text{cov}(\widehat{\boldsymbol{\beta}}_{\text{RI}})$ by

$$\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}}_{\text{RI}}) = \boldsymbol{S}_{xx}^{-1}\left[\frac{\widehat{\mathcal{S}}_1(Y^{\text{obs}}\boldsymbol{X})}{n_1} + \frac{\widehat{\mathcal{S}}_0(Y^{\text{obs}}\boldsymbol{X})}{n_0}\right]\boldsymbol{S}_{xx}^{-1},$$

recalling the definitions of the sample covariance operators $\widehat{\mathcal{S}}_1$ and $\widehat{\mathcal{S}}_0$ introduced in (2). Similar to Neyman (1923), this implicitly assumes $\mathcal{S}(\tau\boldsymbol{X}) = \boldsymbol{0}$. Under the assumption that $\varepsilon_i = 0$ for all units (i.e., no idiosyncratic variation whatsoever), we can instead use $\mathcal{S}(\widehat{\tau}\boldsymbol{X})$ with $\widehat{\tau} = \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_{\text{RI}}$ as a plug-in estimate for $\mathcal{S}(\tau\boldsymbol{X})$. This yields tighter standard errors based on the diagonal elements of the covariance matrix.

Finite Population Asymptotic Analysis. Theorem 2 holds for any finite sample. To obtain confidence intervals and to conduct hypothesis testing as we describe below, we need to prove further

that $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and covariance $\mathrm{cov}(\widehat{\boldsymbol{\beta}}_{\mathrm{RI}})$. Finite population asymptotic analysis, however, has a slightly different flavor from the usual super population approach. Formally, the finite asymptotic scheme embeds the finite population $\{(\boldsymbol{X}_i, Y_i(1), Y_i(0), T_i)\}_{i=1}^n$ with size n into a hypothetical sequence of finite populations with sizes approaching infinity. This effectively assumes that all the finite population quantities, for example, $\boldsymbol{S}_{xx}$ and $\boldsymbol{\beta}$, depend on n, although they are fixed numbers for a given finite population. Moreover, the sample quantities such as $\widehat{\boldsymbol{S}}_{x1}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ depend on n as well, and are random quantities due to the randomization of $\boldsymbol{T}$. For notational simplicity, we drop the index n for all these quantities. Importantly, we must impose some regularity conditions on the hypothetical sequence of finite populations. Throughout the article, we invoke the following conditions for asymptotic analysis, which are required for a form of the finite population central limit theorem discussed in Li and Ding (2017, Theorem 5).

Condition 1. (i) Stable treatment proportions: $p_1 = n_1/n$ and $p_0 = n_0/n$ have positive limiting values; (ii) Stable means, variances, and covariances: the finite population means, variances, and covariances of the covariates and potential outcomes have finite and nonzero limiting values; (iii) both $\boldsymbol{S}_{xx}$ and its limit have full-rank K; (iv) there are no individual extreme values in the limit: $\max_{1\le i\le n} ||V_i - \bar{V}||_2^2/n \to 0$, for $V_i = X_{ki}, Y_i(z), Y_i(z)X_{ki}, X_{ki}X_{k'i}, Y_i(z)X_{ki}X_{k'i}$ with $1 \le k, k' \le K$, and $z = 0, 1$.

Condition parts (i) and (ii) are natural. Part (iii) is a basic requirement for asymptotic analysis of quantities depending on $\boldsymbol{S}_{xx}^{-1}$. The condition on the limit is of particular interest. Having a nonsingular limiting covariance matrix essentially means that there cannot be too many units with extreme leverage on any of the regression coefficients (see Huber 1973). For example, for a binary covariate X_{ki}, the numbers of units with $X_{ki} = 1$ and $X_{ki} = 0$ must both go to infinity. If this did not hold, the limit of $\boldsymbol{S}_{xx}$ would not have full-rank K as the kth row and column would all be driven to 0. Part (iv) controls the tails; it holds if V has more than two moments (Li and Ding 2017). In particular, (iv) holds automatically for bounded covariates and outcomes. For a more technical discussion of finite population causal inference, see Ding (2014), Aronow, Green, and Lee (2014), and Middleton and Aronow (2015); for regularity conditions of the finite population central limit theorems, see Hájek (1960) and Lehmann (1998). A recent review is Li and Ding (2017).

Under these conditions, we can extend Theorem 2 to a sequence of finite populations and obtain a limiting distribution as follows:

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{RI}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \lim_{n\to\infty} \boldsymbol{S}_{xx}^{-1}\left[p_1^{-1}\mathcal{S}\{Y(1)\boldsymbol{X}\} + p_0^{-1}\mathcal{S}\{Y(0)\boldsymbol{X}\} - \mathcal{S}(\tau\boldsymbol{X})\right]\boldsymbol{S}_{xx}^{-1}\right). \quad (7)$$

As a result, we can state that $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ is approximately normal with mean $\boldsymbol{\beta}$ and covariance matrix (6), which allows us to construct confidence intervals and hypothesis tests. In our theory below, we use this informal statement instead of (7) to avoid notational complexity.

3.2. Regression with Treatment-Covariate Interactions

The results from randomization inference can shed light on the familiar case of linear regression with treatment-covariate interactions. This classical approach assumes the model

$$Y_i^{\mathrm{obs}} = \boldsymbol{X}_i^{\top}\boldsymbol{\gamma} + T_i\boldsymbol{X}_i^{\top}\boldsymbol{\beta} + u_i, \qquad (i = 1, \ldots, n), \quad (8)$$

where $\{u_i\}_{i=1}^n$ are errors implicitly assumed to induce the randomness, and where $\boldsymbol{\beta}$ models systematic treatment effect variation, as in (3). Departing from much of the previous literature (e.g., Cox 1984; Berrington de González and Cox 2007; Crump et al. 2008), we study the properties of the least-square estimator under complete randomization, without assuming that model (8) is correctly specified. In particular, we do not assume any iid sampling; the assignment mechanism drives the distribution of the OLS estimator.

Theorem 3. The OLS estimator for $\boldsymbol{\beta}$ from fitting model (8) can be rewritten as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \widehat{\boldsymbol{S}}_{xx,1}^{-1}\widehat{\boldsymbol{S}}_{x1} - \widehat{\boldsymbol{S}}_{xx,0}^{-1}\widehat{\boldsymbol{S}}_{x0},$$

where

$$\widehat{\boldsymbol{S}}_{xx,t} = \frac{1}{n_t}\sum_{i=1}^{n} I_{(T_i=t)}\boldsymbol{X}_i\boldsymbol{X}_i^{\top}, \qquad (t = 0, 1).$$

Over all possible randomizations of $\boldsymbol{T}$, $\widehat{\boldsymbol{S}}_{xx,1}^{-1}\widehat{\boldsymbol{S}}_{x1}$ and $\widehat{\boldsymbol{S}}_{xx,0}^{-1}\widehat{\boldsymbol{S}}_{x0}$ are consistent estimates of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_0$, respectively; $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ therefore follows an asymptotic normal distribution with mean $\boldsymbol{\beta}$ and covariance matrix

$$\mathrm{cov}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \boldsymbol{S}_{xx}^{-1}\left[\frac{\mathcal{S}\{e(1)\boldsymbol{X}\}}{n_1} + \frac{\mathcal{S}\{e(0)\boldsymbol{X}\}}{n_0} - \frac{\mathcal{S}(\varepsilon\boldsymbol{X})}{n}\right]\boldsymbol{S}_{xx}^{-1} \quad (9)$$

with $e_i(1)$, $e_i(0)$, and ε_i as defined after (5).

This estimate is simply the difference between $\widehat{\boldsymbol{\gamma}}_{1,\mathrm{OLS}} = \widehat{\boldsymbol{S}}_{xx,1}^{-1}\widehat{\boldsymbol{S}}_{x1}$ and $\widehat{\boldsymbol{\gamma}}_{0,\mathrm{OLS}} = \widehat{\boldsymbol{S}}_{xx,0}^{-1}\widehat{\boldsymbol{S}}_{x0}$, two OLS regressions run separately on each treatment arm. The (asymptotic) covariance formula (9) is different from (6), with $\{Y(1), Y(0)\}$ replaced by $\{e(1), e(0)\}$. For treated units, define residual $\widehat{e}_i = Y_i^{\mathrm{obs}} - \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\gamma}}_{1,\mathrm{OLS}}$, and for control units, define residual $\widehat{e}_i = Y_i^{\mathrm{obs}} - \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\gamma}}_{0,\mathrm{OLS}}$. We can drop the unidentifiable term $\mathcal{S}(\varepsilon\boldsymbol{X})$, estimate $\mathcal{S}\{e(1)\boldsymbol{X}\}$ and $\mathcal{S}\{e(0)\boldsymbol{X}\}$ by their sample analogs, and conservatively estimate the asymptotic covariance matrix (9) by

$$\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \widehat{\boldsymbol{S}}_{xx,1}^{-1}\left[\frac{\widehat{\mathcal{S}}_1(\widehat{e}\boldsymbol{X})}{n_1}\right]\widehat{\boldsymbol{S}}_{xx,1}^{-1} + \widehat{\boldsymbol{S}}_{xx,0}^{-1}\left[\frac{\widehat{\mathcal{S}}_0(\widehat{e}\boldsymbol{X})}{n_0}\right]\widehat{\boldsymbol{S}}_{xx,0}^{-1}.$$

This form of the sandwich variance estimator has the same probability limit as the Huber–White covariance estimator for linear model (8) (Huber 1967; White 1980; Angrist and Pischke 2008; Lin 2013).

Importantly, $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ are quite similar in form. In particular, $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ uses the true $\boldsymbol{S}_{xx}$ while $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ separately estimates the covariance matrix for each treatment arm, $\widehat{\boldsymbol{S}}_{xx,0}$ and $\widehat{\boldsymbol{S}}_{xx,1}$. The latter is effectively a ratio estimator. Although this introduces some small bias (on the order of $1/n$), using the estimated $\widehat{\boldsymbol{S}}_{xx,t}$ rather than true $\boldsymbol{S}_{xx}$ can often lead to gains in precision, especially when covariates are strongly correlated with the potential outcomes. In particular, the OLS estimator, by separately

estimating the (known) $\boldsymbol{S}_{xx}$ matrix for each treatment arm, can account for random imbalances in the covariates in both arms. For related discussion, see Cochran (1977) on ratio estimators in surveys.

The RI estimator, by comparison, has no adjustment whatsoever, and so cannot account for such random covariate imbalances. However, in Section 3.4 and in the supplementary materials, we introduce a different form of adjustment that uses covariates to make the estimates of the $\boldsymbol{S}_{xt}$ more precise. Depending on the structure of covariates, this estimator could be better or worse than OLS adjustment; we leave a thorough investigation of these trade-offs for future work.

Regardless, we again emphasize that we do *not* rely on classical OLS assumptions to justify the OLS estimator here. Rather, randomization (with some mild regularity conditions for the finite sample asymptotics) justifies our results.

3.3. Omnibus Test for Systematic Variation

Finally, we can use these results to develop an omnibus test for the presence of any systematic treatment effect variation. The null hypothesis of no treatment effect variation explained by the observed covariates can be characterized by

$$H_0(\boldsymbol{X}) : \boldsymbol{\beta}_1 = 0,$$

where $\boldsymbol{\beta}_1$ contains all the components of $\boldsymbol{\beta}$ except the first component corresponding to the intercept. Under $H_0(\boldsymbol{X})$, the individual treatment effects have no linear dependence on $\boldsymbol{X}$.

We then construct a Wald-type test for $H_0(\boldsymbol{X})$ using an estimator $\widehat{\boldsymbol{\beta}}$ and its covariance estimator $\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}})$; it could be $\widehat{\boldsymbol{\beta}}_{\text{RI}}$ or $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$. Let $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}}_1)$ denote the subvector of $\widehat{\boldsymbol{\beta}}$ and submatrix of $\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}})$, corresponding to the nonintercept coordinates of $\boldsymbol{X}$. We reject when

$$\widehat{\boldsymbol{\beta}}_1^{\top}\widehat{\text{cov}}^{-1}(\widehat{\boldsymbol{\beta}}_1)\widehat{\boldsymbol{\beta}}_1 > q_{K-1}(1-\alpha), \tag{10}$$

where $q_{K-1}(1-\alpha)$ is the $1-\alpha$ quantile of the χ^2 random variable with degrees of freedom $K-1$.

The test in (10) is nearly identical to the test proposed by Crump et al. (2008). They relax the parametric assumption by taking a "sieve estimator" approach, namely, by using a quadratic form of the regression function, which allows for more flexible marginal distributions. Our approach differs in that we avoid modeling the marginal distributions entirely. If desired, we can add polynomials of $\boldsymbol{X}$ (or other basis functions) into the model for δ to allow for more flexible systematic treatment effect variation, which could enhance power or model more complex relationships between the $\boldsymbol{X}$ and treatment impact.

3.4. Additional Considerations

In the supplementary material, we describe two additional points about systematic treatment effect variation that we briefly address here. First, as mentioned above, we can use model-assisted estimation to improve the randomization-based estimator. In particular, improving estimation of $\widehat{\boldsymbol{S}}_{xt}$ directly improves $\widehat{\boldsymbol{\beta}}_{\text{RI}}$, as the $\widehat{\boldsymbol{S}}_{xt}$ are the only random components. Thus, if we replace the standard sample estimator, $\widehat{\boldsymbol{S}}_{xt}$, by a more efficient, model-assisted estimator, as in survey sampling (Cochran 1977; Särndal, Swensson, and Wretman 2003), we can achieve meaningful precision gains in practice. More importantly, this setup allows researchers to assess systematic variation across one set of covariates while adjusting for another set.

Second, under the assumption of no idiosyncratic variation (i.e., $\varepsilon_i = 0$ for all i), we can obtain exact inference for $\boldsymbol{\beta}$ by inverting a sequence of randomization-based tests. This complements previous work on randomization-based tests for the presence of idiosyncratic treatment effect variation (Ding, Feller, and Miratrix 2016).

4. Idiosyncratic Treatment Effect Variation for ITT

After characterizing the systematic component of treatment effect variation, we now turn to characterizing the idiosyncratic component. Since this quantity is inherently unidentifiable, we propose sharp bounds on this component and a framework for sensitivity analysis. We then leverage these results to bound an R^2-like measure of the treatment effect variation explained by covariates.

4.1. Bounds

We first define the main quantities of interest:

$$S_{\tau\tau} = \frac{1}{n}\sum_{i=1}^{n}(\tau_i - \tau)^2, \quad S_{\delta\delta} = \frac{1}{n}\sum_{i=1}^{n}(\delta_i - \tau)^2,$$
$$S_{\varepsilon\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2,$$

with δ_i and ε_i defined as in (3). Then $S_{\tau\tau} = S_{\delta\delta} + S_{\varepsilon\varepsilon}$. We can immediately estimate $S_{\delta\delta}$ via the sample variance of $\{\widehat{\delta}_i = \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}\}_{i=1}^n$, where $\widehat{\boldsymbol{\beta}}$ is a consistent estimator, for example, $\widehat{\boldsymbol{\beta}}_{\text{RI}}$ or $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$. However, the idiosyncratic variance, $S_{\varepsilon\varepsilon}$, is inherently unidentifiable because it depends on the joint distribution of potential outcomes.

We can, however, derive sharp bounds for $S_{\varepsilon\varepsilon}$. Let $F_1(y)$ and $F_0(y)$ be the empirical cumulative distribution functions of $\{e_i(1)\}_{i=1}^n$ and $\{e_i(0)\}_{i=1}^n$. Let $F_1^{-1}(u)$ and $F_0^{-1}(u)$ be the corresponding empirical quantile functions, with $F^{-1}(u) = \inf\{x : F(x) \geq u\}$. Below we denote $e(t)$ as a random variable taking equal probabilities on n values of $\{e_i(t)\}_{i=1}^n$. Based on the Fréchet–Hoeffding bounds (Hoeffding 1941; Fréchet 1951; Nelsen 2007), we can bound $S_{\varepsilon\varepsilon}$ as follows.

Theorem 4. $S_{\varepsilon\varepsilon}$ has sharp bounds $\underline{S}_{\varepsilon\varepsilon} \leq S_{\varepsilon\varepsilon} \leq \overline{S}_{\varepsilon\varepsilon}$, where

$$\underline{S}_{\varepsilon\varepsilon} = \int_0^1 \{F_1^{-1}(u) - F_0^{-1}(u)\}^2 du,$$
$$\overline{S}_{\varepsilon\varepsilon} = \int_0^1 \{F_1^{-1}(u) - F_0^{-1}(1-u)\}^2 du.$$

The lower and upper bounds are attainable when $e(1)$ and $e(0)$ have the same ranks and opposite ranks, respectively.

The lower bound of $S_{\varepsilon\varepsilon}$ corresponds to a rank-preserving relationship between $e(1)$ and $e(0)$, and the upper bound of $S_{\varepsilon\varepsilon}$ corresponds to an anti-rank-preserving relationship between $e(1)$ and $e(0)$. Equivalently, they correspond to the cases where

the Spearman rank correlation coefficients between $e(1)$ and $e(0)$ are $+1$ and -1.

In practice, we can often sharpen these bounds because we are unlikely to have negatively associated potential outcomes after adjusting for covariates. If we assume a nonnegative correlation between $e(1)$ and $e(0)$, we have the following corollary.

Corollary 1. If the correlation between $e(1)$ and $e(0)$ is nonnegative, then the bounds for $S_{\varepsilon\varepsilon}$ become $\underline{S}_{\varepsilon\varepsilon} \leq S_{\varepsilon\varepsilon} \leq V_1 + V_0$, where V_t is the variance of $e(t)$ for $t = 0, 1$.

We can consistently estimate each quantity: $S_{\delta\delta}$ by the sample variance of $\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}$, $F_{e1}(y)$ and $F_{e0}(y)$ by $\widehat{F}_1(y)$ and $\widehat{F}_0(y)$, the empirical cumulative distribution functions of the residuals $\widehat{e}_i$ under treatment and control, and V_1 and V_0 by the variances of $\widehat{e}(1)$ and $\widehat{e}(0)$.

Variance of the Overall ITT Estimator. We can use these results to obtain sharper bounds on the variance of Neyman's (1923) estimate of overall ITT, $\widehat{\tau} = n_1^{-1}\sum_{i=1}^{n} T_i Y_i^{\text{obs}} - n_0^{-1}\sum_{i=1}^{n}(1 - T_i)Y_i^{\text{obs}}$, extending previous work by Heckman, Smith, and Clements (1997) and Aronow, Green, and Lee (2014). See also Fogarty (2016). Applying the results in Section 2 for scalar outcomes, we have the following variance for the difference-in-means estimator,

$$\text{var}(\widehat{\tau}) = \frac{S_{11}}{n_1} + \frac{S_{00}}{n_0} - \left(\frac{S_{\delta\delta}}{n} + \frac{S_{\varepsilon\varepsilon}}{n}\right),$$

where $S_{\tau\tau} = S_{\delta\delta} + S_{\varepsilon\varepsilon}$. As we discuss above, Neyman (1923) proposed a lower bound for the overall $\text{var}(\widehat{\tau})$ under the assumption of a constant treatment effect, $S_{\tau\tau} = 0$. More recently, Aronow, Green, and Lee (2014) instead proposed to bound $S_{\tau\tau}$ via Fréchet–Hoeffding bounds. We can modestly improve these results by applying Fréchet–Hoeffding bounds for $S_{\varepsilon\varepsilon}$ alone rather than for $S_{\tau\tau} = S_{\delta\delta} + S_{\varepsilon\varepsilon}$. So long as $S_{\delta\delta} > 0$, this yields strictly tighter bounds on $\text{var}(\widehat{\tau})$ than the corresponding bounds that do not incorporate covariate information. In turn, this gives a tighter estimate of the standard error for the same difference-in-means estimator, $\widehat{\tau}$.

A Variance Ratio Test. Finally, while the relationship between $e(0)$ and $e(1)$ is inherently unidentifiable, there is some information in the data about the relationship between ε_i, the individual-level idiosyncratic treatment effect, and $Y_i(0)$, the control potential outcome. In particular, Raudenbush and Bloom (2015) noted that if the variance of the treatment potential outcomes is smaller than the variance of the control potential outcomes, then the treatment effect must be negatively associated with the control potential outcomes. In the supplementary material, we extend this result to incorporate covariates and propose a formal test.

4.2. Sensitivity Analysis

Going beyond worst-case bounds, we can assess the sensitivity of our estimate of $S_{\varepsilon\varepsilon}$ to different assumptions of the dependence between potential outcomes. Using the probability integral transformation, we represent the residual potential outcomes as

$$e(1) = F_1^{-1}(U_1), \quad e(0) = F_0^{-1}(U_0), \quad U_1, U_0 \sim \text{Uniform}(0, 1).$$

Therefore, the dependence of the potential outcomes is determined by the dependence of the uniform random variables U_1 and U_0, which are the standardized ranks of the potential outcomes. When $U_1 = U_0$, $S_{\varepsilon\varepsilon}$ attains the lower bound $\underline{S}_{\varepsilon\varepsilon}$; when $U_1 = 1 - U_0$, $S_{\varepsilon\varepsilon}$ attains the upper bound $\overline{S}_{\varepsilon\varepsilon}$; when $U_1 \perp\!\!\!\perp U_0$, $S_{\varepsilon\varepsilon}$ attains the improved upper bound $V_1 + V_0$.

Rather than simply examine extreme scenarios of $S_{\varepsilon\varepsilon}$, we can instead represent U_1 as a mixture of U_0 and another independent uniform random variable V_0 :

$$U_1 \sim \rho U_0 + (1 - \rho)V_0, \quad U_0, V_0 \overset{\text{iid}}{\sim} \text{Uniform}(0, 1), \tag{11}$$

which the sensitivity parameter ρ captures the association between U_1 and U_0. An immediate interpretation of ρ is the proportion of rank preserved units, with the other $1 - \rho$ as the proportion of units with independent treatment and control residual outcomes. When $\rho = 0$, $U_1 \perp\!\!\!\perp U_0$, and the residual potential outcomes are independent; when $\rho = 1$, $U_1 = U_0$, and the residual potential outcomes have the same ranks. The values between $(0, 1)$ correspond to positive rank correlation but not full-rank preservation. Note that the representation of the joint distribution is not unique, because we can choose any copula as a joint distribution of (U_1, U_0) (Nelsen 2007). We choose the above representation and notation ρ for the following theorem.

Theorem 5. If Equation (11) holds, then ρ is Spearman's rank correlation coefficient between $e(1)$ and $e(0)$. Furthermore, $S_{\varepsilon\varepsilon}$ is a linear function of ρ:

$$S_{\varepsilon\varepsilon}(\rho) = \rho\underline{S}_{\varepsilon\varepsilon} + (1 - \rho)(V_1 + V_0).$$

We cannot extract any information about ρ from the data. We therefore treat ρ as a sensitivity parameter, choose a plausible range of ρ, and obtain corresponding values for $S_{\varepsilon\varepsilon}$.

4.3. Fraction of Treatment Effect Variation Explained

A natural question is the relative magnitudes of $S_{\delta\delta}$ and $S_{\varepsilon\varepsilon}$ (Djebbari and Smith 2008). Continuing the regression analogy, this is an R^2-like measure for the proportion of total treatment effect variation explained by the systematic component:

$$R_\tau^2 = \frac{S_{\delta\delta}}{S_{\tau\tau}} = \frac{S_{\delta\delta}}{S_{\delta\delta} + S_{\varepsilon\varepsilon}},$$

which is the ratio between the finite population variances of δ and τ. As above, we can directly estimate $S_{\delta\delta}$ but must bound $S_{\varepsilon\varepsilon}$. Applying Theorem 4, we obtain the following bounds on R_τ^2.

Corollary 2. The sharp bounds on R_τ^2 are

$$\frac{S_{\delta\delta}}{S_{\delta\delta} + \overline{S}_{\varepsilon\varepsilon}} \leq R_\tau^2 \leq \frac{S_{\delta\delta}}{S_{\delta\delta} + \underline{S}_{\varepsilon\varepsilon}}.$$

If we further assume that the correlation between $e(1)$ and $e(0)$ is nonnegative, the sharp bounds on R_τ^2 are

$$\frac{S_{\delta\delta}}{S_{\delta\delta} + V_1 + V_0} \leq R_\tau^2 \leq \frac{S_{\delta\delta}}{S_{\delta\delta} + \underline{S}_{\varepsilon\varepsilon}}.$$

We estimate these bounds via plug-in estimates. Note that Djebbari and Smith (2008) explored a similar quantity by using a permutation approach to approximate the

Fréchet–Hoeffding upper and lower bounds. Finally, we can use the sensitivity results for $S_{\varepsilon\varepsilon}$, with values of $\rho \in [0, 1]$:

$$R^2_\tau(\rho) = \frac{S_{\delta\delta}}{S_{\delta\delta} + S_{\varepsilon\varepsilon}(\rho)}.$$

5. Noncompliance

5.1. Setup

We now extend our results to allow for noncompliance. Let T be the indicator of treatment assigned, D be the indicator of treatment received, Y be outcome of interest, and $\boldsymbol{X}$ be pretreatment covariates. Under the Stable Unit Treatment Value Assumption, we define $D_i(t)$ and $Y_i(t)$ as the potential outcomes for unit i under treatment assignment t. Following Angrist, Imbens, and Rubin (1996) and Frangakis and Rubin (2002), we can classify units into four compliance types based on the joint values of $D_i(1)$ and $D_i(0)$:

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

Denote n_u and π_u as the number and proportion of compliance types π_u of stratum $U = u$ for $u = a, n, c, d$.

Throughout our discussion, we invoke the following assumptions which are commonly used for analyzing randomized experiments with noncompliance.

Assumption 1. (i) Monotonicity: $D_i(1) \geq D_i(0)$; (ii) Exclusion restrictions for Always Takers and Never Takers: $Y_i(1) = Y_i(0)$ for all units with $D_i(1) = D_i(0)$; (iii) Strong instrument: $\pi_c > C_0 > 0$, where C_0 is a positive constant independent of the sample size.

Monotonicity rules out the existence of Defiers, that is, $\pi_d = 0$. Under monotonicity, we can estimate the proportion π_u using the observed counts of units classified by T and D: let $n_{td} = \#\{i : T_i = t, D_i = d\}$, and then $\widehat{\pi}_n = n_{10}/n_1, \widehat{\pi}_a = n_{01}/n_0$, and $\widehat{\pi}_c = n_{11}/n_1 - n_{01}/n_0$. The exclusion restrictions assume that treatment assignment has no effect on the outcome for Always Takers and Never Takers. As a result, treatment effect variation is trivially zero for Always Takers and Never Takers. Note that this is the unit-level exclusion restriction imposed in Angrist, Imbens, and Rubin (1996). This can be relaxed in other settings, for example, we could assume the impact of randomization for these groups is zero on average (see Imbens and Rubin 2015). Finally, to avoid technical complexity, we rule out the weak instrument case (Bound, Jaeger, and Baker 1995; Staiger and Stock 1997), that is, π_c is within a small neighborhood of 0 with radius shrinking to 0.

We are interested in treatment effect variation among Compliers, which motivates the following decomposition:

$$\tau_i = Y_i(1) - Y_i(0) = \begin{cases} 0, & \text{if } U_i = a \text{ or } n, \\ \boldsymbol{X}_i^\top \boldsymbol{\beta}_c + \varepsilon_i, & \text{if } U_i = c, \end{cases} \quad (12)$$

where $\boldsymbol{\beta}_c$ is the regression coefficient of τ_i on $\boldsymbol{X}_i$ among Compliers, analogous to (3).

5.2. Systematic Treatment Effect Variation Among Compliers

5.2.1. Randomization Inference

We now extend the results of Section 3 to estimate systematic treatment effect variation among Compliers. Define

$$\boldsymbol{S}_{xx,u} = \frac{1}{n_u}\sum_{i=1}^n I_{(U_i=u)}\boldsymbol{X}_i\boldsymbol{X}_i^\top, \quad \boldsymbol{S}_{xt,u} = \frac{1}{n_u}\sum_{i=1}^n I_{(U_i=u)}Y_i(t)\boldsymbol{X}_i, \quad (t = 0, 1; u = a, c, n).$$

Then, analogous to (5),

$$\boldsymbol{\beta}_c = \boldsymbol{S}_{xx,c}^{-1}(\boldsymbol{S}_{x1,c} - \boldsymbol{S}_{x0,c}) = \boldsymbol{S}_{xx,c}^{-1}\boldsymbol{S}_{x1,c} - \boldsymbol{S}_{xx,c}^{-1}\boldsymbol{S}_{x0,c} \equiv \boldsymbol{\gamma}_{1c} - \boldsymbol{\gamma}_{0c}, \quad (13)$$

where

$$\boldsymbol{\gamma}_{1c} = \boldsymbol{S}_{xx,c}^{-1}\boldsymbol{S}_{x1,c}, \quad \boldsymbol{\gamma}_{0c} = \boldsymbol{S}_{xx,c}^{-1}\boldsymbol{S}_{x0,c}$$

are the linear regression coefficients of $Y(1)$ and $Y(0)$ on covariates among Compliers.

Unlike in the ITT case, we cannot estimate these quantities directly. Instead, following standard results from noncompliance (e.g., Angrist, Imbens, and Rubin 1996; Abadie 2003; Angrist and Pischke 2008), we use estimates from observed subgroups to estimate the desired quantities of interest. Define sample moments:

$$\widehat{\boldsymbol{S}}_{xx,td} = \frac{1}{n_t}\sum_{i=1}^n I_{(T_i=t)}I_{(D_i=d)}\boldsymbol{X}_i\boldsymbol{X}_i^\top,$$
$$\widehat{\boldsymbol{S}}_{xt,td} = \frac{1}{n_t}\sum_{i=1}^n I_{(T_i=t)}I_{(D_i=d)}Y_i^{\text{obs}}\boldsymbol{X}_i \quad (t, d = 0, 1). \quad (14)$$

The following theorem connects these quantities with the finite population quantities in (13).

Theorem 6. Over all possible randomizations of a completely randomized experiment, both $\widehat{\boldsymbol{S}}_{xx}(1) = \widehat{\boldsymbol{S}}_{xx,11} - \widehat{\boldsymbol{S}}_{xx,01}$ and $\widehat{\boldsymbol{S}}_{xx}(0) = \widehat{\boldsymbol{S}}_{xx,00} - \widehat{\boldsymbol{S}}_{xx,10}$ are unbiased for $\pi_c\boldsymbol{S}_{xx,c}$, and

$$E(\widehat{\boldsymbol{S}}_{x1,11} - \widehat{\boldsymbol{S}}_{x0,01}) = \pi_c\boldsymbol{S}_{x1,c}, \quad E(\widehat{\boldsymbol{S}}_{x0,00} - \widehat{\boldsymbol{S}}_{x1,10}) = \pi_c\boldsymbol{S}_{x0,c}. \quad (15)$$

This theorem shows that we can obtain unbiased estimates for all terms in (13). The following corollary shows that we can then obtain consistent estimates for $\boldsymbol{\gamma}_{1c}$, $\boldsymbol{\gamma}_{0c}$, and $\boldsymbol{\beta}_c$, recalling that in the asymptotic analysis, we need to embed $\{(\boldsymbol{X}_i, Y_i(1), Y_i(0), D_i(1), D_i(0), T_i)\}_{i=1}^n$ into a hypothetical sequence of finite populations under Condition 1 and the following Condition 2.

Condition 2. Both $\boldsymbol{S}_{xx,c}$ and its limit have full-rank K.

Condition 2 holds if and only if any linear combination of $\boldsymbol{X}$, $\boldsymbol{l}^\top\boldsymbol{X}$ with $\boldsymbol{l} \neq \boldsymbol{0}$, has positive finite population variance among Compliers. Condition 2 is effectively the finite population version of ruling out weak instruments in the two-stage least-square estimate with treatment-covariate interactions (e.g., Angrist and Pischke 2008).

Corollary 3. $\widehat{\boldsymbol{\gamma}}_{1c,\mathrm{RI}} = \widehat{\boldsymbol{S}}_{xx}^{-1}(1)(\widehat{\boldsymbol{S}}_{x1,11} - \widehat{\boldsymbol{S}}_{x0,01})$ and $\widehat{\boldsymbol{\gamma}}_{0c,\mathrm{RI}} = \widehat{\boldsymbol{S}}_{xx}^{-1}(0)(\widehat{\boldsymbol{S}}_{x0,00} - \widehat{\boldsymbol{S}}_{x1,10})$ are consistent for $\boldsymbol{\gamma}_{1c}$ and $\boldsymbol{\gamma}_{0c}$. Furthermore, $\widehat{\boldsymbol{\beta}}_{c,\mathrm{RI}} = \widehat{\boldsymbol{\gamma}}_{1c,\mathrm{RI}} - \widehat{\boldsymbol{\gamma}}_{0c,\mathrm{RI}}$ is consistent for $\boldsymbol{\beta}_c$ and follows an asymptotic normal distribution with covariance matrix

$$\mathrm{cov}(\widehat{\boldsymbol{\beta}}_{c,\mathrm{RI}}) = (\pi_c \boldsymbol{S}_{xx,c})^{-1}\left[\frac{\mathcal{S}\{e'(1)\boldsymbol{X}\}}{n_1} + \frac{\mathcal{S}\{e'(0)\boldsymbol{X}\}}{n_0} - \frac{\mathcal{S}(\varepsilon\boldsymbol{X})}{n}\right](\pi_c \boldsymbol{S}_{xx,c})^{-1}, \tag{16}$$

where we define the residual potential outcomes to be:

$$e_i'(1) = \begin{cases} Y_i(1) - \boldsymbol{X}_i^{\top}\boldsymbol{\gamma}_{1c}, \\ Y_i(1) - \boldsymbol{X}_i^{\top}\boldsymbol{\gamma}_{0c}, \\ Y_i(1) - \boldsymbol{X}_i^{\top}\boldsymbol{\gamma}_{1c}, \end{cases} \quad e_i'(0) = \begin{cases} Y_i(0) - \boldsymbol{X}_i^{\top}\boldsymbol{\gamma}_{1c}, & U_i = a, \\ Y_i(0) - \boldsymbol{X}_i^{\top}\boldsymbol{\gamma}_{0c}, & U_i = n, \\ Y_i(0) - \boldsymbol{X}_i^{\top}\boldsymbol{\gamma}_{0c}, & U_i = c. \end{cases} \tag{17}$$

The idiosyncratic variation is $\varepsilon_i = e_i'(1) - e_i'(0)$ for unit i, with $\varepsilon_i = 0$ for Never Takers and Always Takers, and with ε_i for Compliers as in (12). The two sets of residuals are not formed from a regression on all units, but instead the population regression on Compliers alone. As in the ITT case, we can estimate $\mathcal{S}\{e'(1)\boldsymbol{X}\}$ and $\mathcal{S}\{e'(0)\boldsymbol{X}\}$ using their sample analogs; $\mathcal{S}(\varepsilon\boldsymbol{X})$, however, is unidentifiable. For units with $D_i = 1$, we define the residual $\widehat{e}_i' = Y_i^{\mathrm{obs}} - \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\gamma}}_{c1,\mathrm{RI}}$, and for units with $D_i = 0$, we define the residual $\widehat{e}_i' = Y_i^{\mathrm{obs}} - \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\gamma}}_{c0,\mathrm{RI}}$. Therefore, we can obtain a conservative estimate for the asymptotic covariance (16) by the following sandwich form:

$$\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}_{c,\mathrm{RI}}) = \widehat{\boldsymbol{S}}_{xx}^{-1}(1)\left[\frac{\widehat{\mathcal{S}}_1(\widehat{e}'\boldsymbol{X})}{n_1}\right]\widehat{\boldsymbol{S}}_{xx}^{-1}(1) + \widehat{\boldsymbol{S}}_{xx}^{-1}(0)\left[\frac{\widehat{\mathcal{S}}_0(\widehat{e}'\boldsymbol{X})}{n_0}\right]\widehat{\boldsymbol{S}}_{xx}^{-1}(0).$$

As with the ITT analog, so long as we have Assumption 1, randomization itself fully justifies the theorem and estimators without relying on a model of the observed outcomes.

5.2.2. Two-Stage Least Squares

We now turn to the standard two-stage least-square (TSLS) setting in econometrics (e.g., Angrist and Pischke 2008). First, we impose a linear regression model with treatment-covariate interactions:

$$Y_i^{\mathrm{obs}} = \boldsymbol{X}_i^{\top}\boldsymbol{\gamma} + D_i\boldsymbol{X}_i^{\top}\boldsymbol{\beta} + u_i \quad (i = 1, \ldots, n).$$

Here, the randomness of the observed outcome comes from the randomness of D_i and u_i. In the language of econometrics, the treatment received is "endogenous," that is, D_i and the error term u_i are assumed to be correlated; we therefore use T_i as an instrument for D_i. The TSLS estimates $(\widehat{\boldsymbol{\gamma}}_{\mathrm{TSLS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{TSLS}})$ are the solutions to the following estimating equations:

$$n^{-1}\sum_{i=1}^{n}\begin{pmatrix}\boldsymbol{X}_i \\ T_i\boldsymbol{X}_i\end{pmatrix}(Y_i^{\mathrm{obs}} - \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\gamma}}_{\mathrm{TSLS}} - D_i\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}_{\mathrm{TSLS}}) = 0. \tag{18}$$

This approach is based on M-estimation, though there are many other ways to formalize the TSLS estimator (e.g., Imbens 2014). The following theorem shows that the fully interacted TSLS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{TSLS}}$ is consistent for $\boldsymbol{\beta}_c$ across randomizations.

Theorem 7. Over all randomizations, the TSLS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{TSLS}}$ follows an asymptotic normal distribution with mean $\boldsymbol{\beta}_c$ and covariance matrix

$$(\pi_c \boldsymbol{S}_{xx,c})^{-1}\left[\frac{\mathcal{S}\{e''(1)\boldsymbol{X}\}}{n_1} + \frac{\mathcal{S}\{e''(0)\boldsymbol{X}\}}{n_0} - \frac{\mathcal{S}(\varepsilon\boldsymbol{X})}{n}\right](\pi_c \boldsymbol{S}_{xx,c})^{-1},$$

where we define the residual potential outcomes to be

$$e_i''(1) = \begin{cases} Y_i(1) - \boldsymbol{X}_i^{\top}(\boldsymbol{\gamma}_{\infty} + \boldsymbol{\beta}_c), & U_i = a, \\ Y_i(1) - \boldsymbol{X}_i^{\top}\boldsymbol{\gamma}_{\infty}, & U_i = n \\ Y_i(1) - \boldsymbol{X}_i^{\top}(\boldsymbol{\gamma}_{\infty} + \boldsymbol{\beta}_c), & U_i = c, \end{cases}$$

$$e_i''(0) = \begin{cases} Y_i(0) - \boldsymbol{X}_i^{\top}(\boldsymbol{\gamma}_{\infty} + \boldsymbol{\beta}_c), & U_i = a, \\ Y_i(0) - \boldsymbol{X}_i^{\top}\boldsymbol{\gamma}_{\infty}, & U_i = n \\ Y_i(0) - \boldsymbol{X}_i^{\top}\boldsymbol{\gamma}_{\infty}, & U_i = c, \end{cases}$$

where $\boldsymbol{\gamma}_{\infty}$ is the probability limit of the TSLS regression coefficient, $\widehat{\boldsymbol{\gamma}}_{\mathrm{TSLS}}$, and the idiosyncratic treatment effect is $\varepsilon_i \equiv e_i''(1) - e_i''(0)$.

For variance estimation, define the residual as $\widehat{e}_i'' = Y_i^{\mathrm{obs}} - \boldsymbol{X}_i^{\top}(\widehat{\boldsymbol{\gamma}}_{\mathrm{TSLS}} + \widehat{\boldsymbol{\beta}}_{\mathrm{TSLS}})$ for units with $D_i = 1$ and $\widehat{e}_i'' = Y_i^{\mathrm{obs}} - \boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\gamma}}_{\mathrm{TSLS}}$ for units with $D_i = 0$. We can then use the following sandwich variance estimator:

$$\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}_{\mathrm{TSLS}}) = \widehat{\boldsymbol{S}}_{xx}^{-1}(1)\left[\frac{\widehat{\mathcal{S}}_1(\widehat{e}''\boldsymbol{X})}{n_1}\right]\widehat{\boldsymbol{S}}_{xx}^{-1}(1) + \widehat{\boldsymbol{S}}_{xx}^{-1}(0)\left[\frac{\widehat{\mathcal{S}}_0(\widehat{e}''\boldsymbol{X})}{n_0}\right]\widehat{\boldsymbol{S}}_{xx}^{-1}(0),$$

which has the same probability limit as the Huber–White covariance estimator for $\widehat{\boldsymbol{\beta}}_{\mathrm{TSLS}}$. Therefore, the randomization itself effectively justifies the use of TSLS for estimating systematic treatment effect variation among Compliers, extending our ITT results.

Finally, while $\widehat{\boldsymbol{\beta}}_{\mathrm{TSLS}}$ is a consistent estimator for $\boldsymbol{\beta}_c$, $\widehat{\boldsymbol{\gamma}}_{\mathrm{TSLS}}$ is not, in general, a consistent estimator for $\boldsymbol{\gamma}_{c0}$, that is, $\boldsymbol{\gamma}_{\infty} \neq \boldsymbol{\gamma}_{c0}$. Instead, $\widehat{\boldsymbol{\gamma}}_{\mathrm{TSLS}}$ converges to $\boldsymbol{\gamma}_{\infty} = \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x0} - \pi_a\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xx,a}\boldsymbol{\beta}_c$. In the special case of one-sided noncompliance (i.e., $\pi_a = 0$), $\boldsymbol{\gamma}_{\infty} = \boldsymbol{\gamma}_0 = \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x0}$, the population OLS regression coefficient, among all Compliers and Never Takers, of $Y(0)$ on covariates.

5.2.3. Omnibus Test for Systematic Treatment Effect Variation Among Compliers

With point estimate $\widehat{\boldsymbol{\beta}}$ and covariance estimate $\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}})$ for $\boldsymbol{\beta}_c$, we can use the same Wald-type χ^2 test as in (10) for the presence of systematic treatment effect variation among Compliers. Here, the estimator can be either randomization-based $\widehat{\boldsymbol{\beta}}_{c,\mathrm{RI}}$ or TSLS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{TSLS}}$; the degrees of freedom are the same, $K - 1$. Unlike in the ITT case, we are not aware of existing tests for systematic treatment effect variation among Compliers.

5.3. Idiosyncratic Treatment Effect Variation with Noncompliance

5.3.1. Bounding Idiosyncratic Variation

We now turn to decomposing the overall treatment effect in the presence of noncompliance. In this setting, we have three sources of treatment effect variation: (i) systematic treatment

effect variation among Compliers, (ii) idiosyncratic treatment effect variation among Compliers, and (iii) treatment effect variation due to noncompliance.

First, recall that total treatment effect variation is $S_{\tau\tau} = \sum_{i=1}^{n}(\tau_i - \tau)^2/n$. We can define a similar quantity among Compliers:

$$S_{\tau\tau,c} = \frac{1}{n_c}\sum_{i=1}^{n} I_{(U_i=c)}(\tau_i - \tau_c)^2.$$

As in Section 4, we can decompose this variation into systematic and idiosyncratic treatment effect variation for Compliers, respectively:

$$S_{\delta\delta,c} = \frac{1}{n_c}\sum_{i=1}^{n} I_{(U_i=c)}(\delta_i - \tau_c)^2, \qquad S_{\varepsilon\varepsilon,c} = \frac{1}{n_c}\sum_{i=1}^{n} I_{(U_i=c)}\varepsilon_i^2.$$

Because treatment effects for Never Takers and Always Takers are zero, there is no treatment effect variation for these units. The component of treatment effect variation due to compliance status is

$$S_{\tau\tau,U} = \sum_{u=c,a,n} \pi_u(\tau_u - \tau)^2.$$

Using $\tau_a = \tau_n = 0$ and $\tau = \pi_c\tau_c$ due to the exclusion restrictions, we have the following theorem summarizing the relationships among the above components.

Theorem 8. $S_{\tau\tau} = \pi_c S_{\tau\tau,c} + S_{\tau\tau,U}$, $S_{\tau\tau,c} = S_{\delta\delta,c} + S_{\varepsilon\varepsilon,c}$, and $S_{\tau\tau,U} = \pi_c(1-\pi_c)\tau_c^2$.

In words, total treatment effect variation has three parts: (i) systematic treatment effect variation among Compliers, $\pi_c S_{\delta\delta,c}$; (ii) idiosyncratic treatment effect variation among Compliers, $\pi_c S_{\varepsilon\varepsilon,c}$; (iii) treatment effect variation due to noncompliance, $S_{\tau\tau,U}$.

As in the ITT case, even though $S_{\varepsilon\varepsilon,c}$ is not identifiable, we can derive bounds in terms of the marginal distributions of the residuals, $\{e_i'(1) = Y_i(1) - \boldsymbol{X}_i^{\top}\boldsymbol{\gamma}_{1c} : U_i = c, i = 1, \ldots, n\}$ and $\{e_i'(0) = Y_i(0) - \boldsymbol{X}_i^{\top}\boldsymbol{\gamma}_{0c} : U_i = c, i = 1, \ldots, n\}$, denoted by $F_{1c}(y)$ and $F_{0c}(y)$, and with marginal variances, V_{1c} and V_{0c}. Once we estimate these quantities, we can plug them in to Theorem 4 and Corollary 1 to get our bounds. As compliance status is only partially observed, we have to estimate these quantities by differencing observed distributions; we defer this and some other technical details to the supplementary material.

5.3.2. Treatment Effect Decomposition

Since there are two sources of variation—covariates and noncompliance—there are three possible R^2-type measures. First, we can measure the treatment effect variation explained by noncompliance alone (i.e., only U):

$$R^2_{\tau,U} = \frac{S_{\tau\tau,U}}{S_{\tau\tau}} = \frac{S_{\tau\tau,U}}{S_{\tau\tau,U} + \pi_c S_{\tau\tau,c}} = \frac{S_{\tau\tau,U}}{S_{\tau\tau,U} + \pi_c S_{\delta\delta,c} + \pi_c S_{\varepsilon\varepsilon,c}}.$$

Second, we can measure the proportion of treatment effect variation among Compliers explained by covariates (i.e., only $\boldsymbol{X}$):

$$R^2_{\tau,c} = \frac{S_{\delta\delta,c}}{S_{\tau\tau,c}} = \frac{S_{\delta\delta,c}}{S_{\delta\delta,c} + S_{\varepsilon\varepsilon,c}}.$$

Third, we can measure the treatment effect variation explained by covariates and noncompliance (i.e., both $\boldsymbol{X}$ and U):

$$R^2_{\tau,U\boldsymbol{X}} = \frac{S_{\tau\tau,U} + \pi_c S_{\delta\delta,c}}{S_{\tau\tau}} = \frac{S_{\tau\tau,U} + \pi_c S_{\delta\delta,c}}{S_{\tau\tau,U} + \pi_c S_{\delta\delta,c} + \pi_c S_{\varepsilon\varepsilon,c}}.$$

For each measure, we can use tailored versions of Corollary 1 to construct bounds, or conduct sensitivity analysis as in Section 4.2, with the sensitivity parameter expressed as the Spearman correlation between the treatment and control potential outcomes among Compliers.

6. Simulation Study

6.1. ITT Estimators

We simulate completely randomized experiments to evaluate the finite sample performance of the tests for systematic treatment effect variation based on $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$, $\widehat{\boldsymbol{\beta}}_{\text{RI}}$, and $\widehat{\boldsymbol{\beta}}^{w}_{\text{RI}}$, the model-assisted version discussed in the supplementary material. Our data generation process is inspired by the Head Start Impact Study (HSIS) study analyzed in the next section. For a given sample size, we first generate four independent covariates (X_1, a standard normal, X_2, a binary covariate with probability 0.5 being 1, X_3, a binary covariate with probability 0.25 being 1, and X_4, a standard normal). The control potential outcomes are then generated from

$$Y_i(0) = 0.3 + 0.2X_{1i} + 0.3X_{2i} - 0.4X_{3i} + 0.8X_{4i} + u_i, \quad u_i \sim \mathcal{N}(0, \sigma^2).$$

We select $\sigma^2 = 0.26$ to make the marginal variance for the control potential outcomes 1; thus we can interpret impacts in "effect size" units. The R^2 of regressing $Y(0)$ onto the covariates is approximately 0.74, due to the "pretest"-like variable X_{4i}. Without X_{4i}, the R^2 is about 0.09.

The treatment effects are $\tau_i = \delta_i + \varepsilon_i$, with (i) either $\delta_i = 0.3$ for all i, or $\delta_i = 0.2 + 0.1X_{1i} + 0.4X_{3i}$; and (ii) either $\varepsilon_i = 0$ for all i, or $\varepsilon_i \sim \mathcal{N}(0, 0.2^2)$. All combinations of these two options give the four cases of (a) no treatment effect variation, (b) only systematic variation, (c) idiosyncratic variation with no systematic variation, and (d) both systematic and idiosyncratic variation. For an α-level test of systematic variation, scenarios (a) and (c) should only reject at rate α, while we would like to see high rejection rates for scenarios (b) and (d). For scenario (d), the R^2_τ is about 0.5; systematic variation explains a good share of the overall variation.

To generate a synthetic dataset, we generated all potential outcomes, randomized units into treatment with probability 0.6, and then calculated the corresponding observed outcomes. We then conducted a test for systematic variation using each of our three estimators. For $\widehat{\boldsymbol{\beta}}_{\text{RI}}$ and $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$, we use X_1, X_2, X_3. For our covariate-adjusted estimator $\widehat{\boldsymbol{\beta}}^{w}_{\text{RI}}$, we also include the fairly predictive X_4 for adjustment.

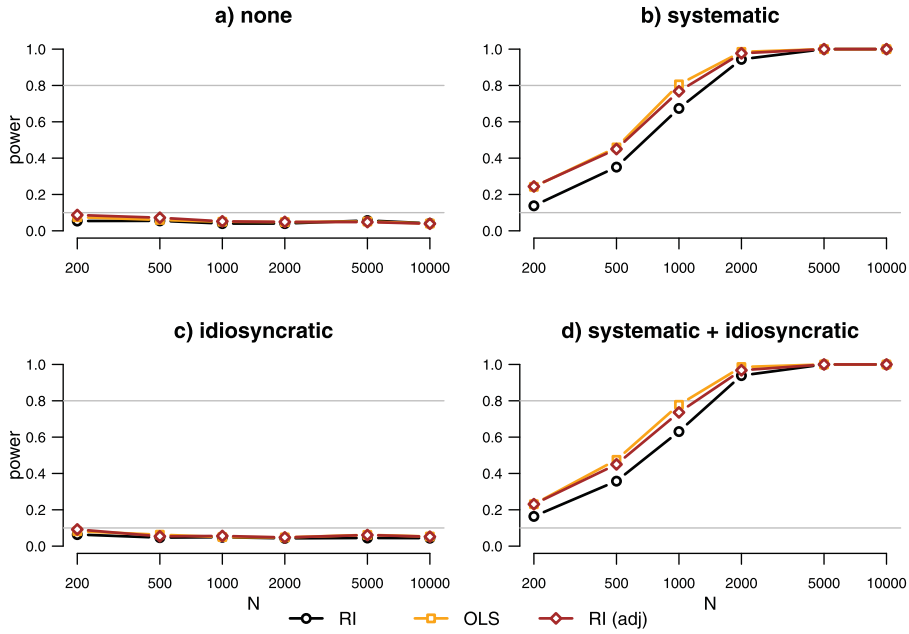


Figure 1. Power of the tests based on $\widehat{\beta}_{\text{RI}}$, $\widehat{\beta}_{\text{OLS}}$, and $\widehat{\beta}^{w}_{\text{RI}}$.

Figure 1 shows the power of these tests, with significance level $\alpha = 0.05$, for different sample sizes. First, all estimators appear asymptotically valid, consistent with the theoretical results. The OLS and adjusted estimators are slightly anti-conservative for small n, however, with rejection rates of around 9%. Second, the OLS estimator appears to have the greatest power in this setting, which is unsurprising since the true data-generating process is a linear model. Finally, covariate adjustment slightly improves the power of the RI estimator. Overall, in the scenarios we consider, we only achieve decent levels of power in large samples, although there seems to be reasonable power for the sample size in the data application, $n = 3586$.

6.2. LATE Estimators

We next simulate completely randomized experiments with noncompliance to evaluate the finite sample performance of the tests for systematic treatment effect variation among Compliers based on $\widehat{\beta}_{c,\text{RI}}$ and $\widehat{\beta}_{\text{TSLS}}$. We first generated a complete dataset as in the ITT case above, and then assigned strata membership to all units with probabilities proportional to their covariates. For Always Takers, we then set $Y_i(0) = Y_i(1)$, and for Never Takers, $Y_i(1) = Y_i(0)$. The overall ITT is now reduced to 0.21 (due to the 0 effects of Never Takers and Always Takers), although the CACE is still approximately 0.3. The proportion of Compliers is approximately 68%.

The Compliers have the systematic and idiosyncratic effects described as above. We tested for the presence of systematic variation for Compliers under the exclusion restrictions. Figure 2 shows the power of these tests for our RI and TSLS estimators. First, in this scenario, the 2SLS and the RI estimators are virtually equivalent; the additional adjustment provided by TSLS does not add significantly to the precision. We see the tests are valid (they even appear conservative) for cases (a) and (c). Power is reduced compared to the ITT simulation; this is reasonable as power is effectively a function of the number of Compliers, with additional uncertainty due to partial information about the identity of Compliers.

7. Application to the Head Start Impact Study

Established in 1965, Head Start is the largest Federal preschool program in the United States, serving nearly 1 million low-income 3- and 4-year-old children each year at a cost of over $7 billion (Administration for Children and Families 2015). Researchers and policymakers have debated Head Start's effectiveness since its inception, with early randomized trials finding limited impacts (e.g., Westinghouse Learning Corporation 1969) and quasi-experimental studies showing much larger effects (e.g., Currie and Thomas 1995). Designed in part to settle this debate, the Head Start Impact Study (HSIS) is a large-scale, nationally representative randomized trial of Head Start first launched in 2002 (Puma et al. 2010). The Congressional mandate for HSIS included two broad questions: (1) the program's *overall* impact, and (2) how impacts *vary* across children and centers. The policy debate has largely focused on this first question; HSIS only found modest average effects on a range of children's cognitive and social-emotional outcomes. However, both the original study and several recent articles argue that these topline results mask important treatment effect variation (e.g., Bloom and Weiland 2014; Bitler, Hoynes, and Domina 2014; Walters 2015; Ding, Feller, and Miratrix 2016; Feller et al. 2016). Understanding such variation is critical both for assessing the program's benefits and costs and for improving the practice and science of early childhood education.

HSIS collected a rich set of covariates about children and their families, including pretest score, child's age, child's race, child's home language, mother's education level, and mother's marital status. At the same time, many potentially important covariates are unavailable. For instance, while families must be low-income to be eligible for Head Start, HSIS does not include information on families' actual income nor other financial details that could be important predictors of program impact.

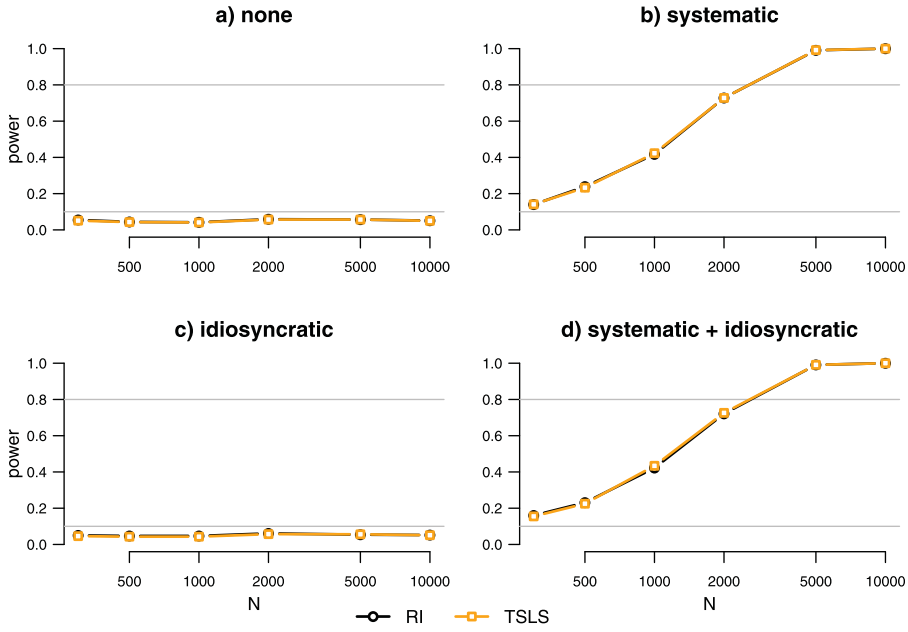


Figure 2. Power of the tests based on $\widehat{\beta}_{c,\mathrm{RI}}$ and $\widehat{\beta}_{\mathrm{TSLS}}$.

In addition, Feller et al. (2016) and others argue that the setting in which a child would otherwise receive care is an important source of impact variation, although this is not directly observable.

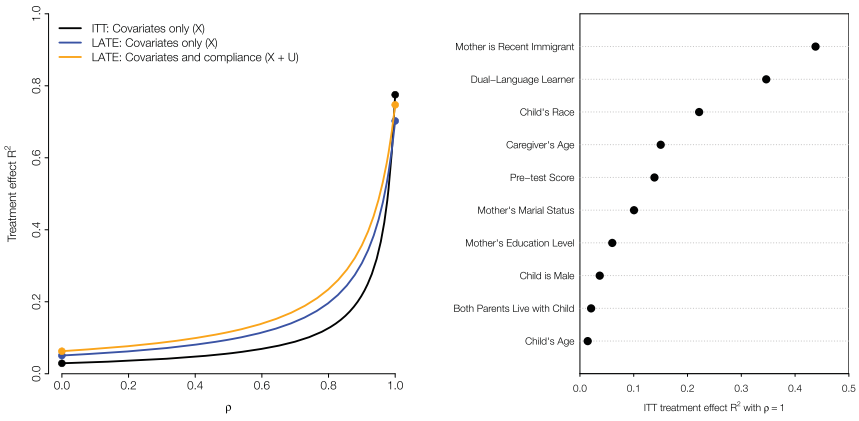
We now use the methods outlined above to assess treatment effect variation in HSIS. The original study included $n = 4400$ total children, with $n_1 = 2644$ in the treatment group and $n_0 = 1796$ in the control group. Following earlier analyses (Ding, Feller, and Miratrix 2016) and to simplify exposition, we restrict our attention to a complete-case subset of the HSIS, with $n_1 = 2238$ in the treatment group and $n_0 = 1348$ in the control group (so $p_1 \approx 0.62$ and $p_0 \approx 0.38$). Our outcome of interest is the Peabody Picture Vocabulary Test (PPVT), a widely used measure of cognitive ability in early childhood. To assess treatment effect variation, we consider the full set of child- and family-level covariates used in the original HSIS analysis of Puma et al. (2010), including those mentioned above. After creating dummy variables for factors (e.g., recoding race), the covariate matrix has 17 columns. See Figure 3(b) for a complete list.

7.1. Decomposing Variation in the ITT Effect

We first explore treatment effect variation for the ITT estimate, beginning with estimating systematic treatment effect variation. We examine three estimators: the randomization-based and OLS estimators discussed in Section 3, $\widehat{\beta}_{\mathrm{RI}}$ and $\widehat{\beta}_{\mathrm{OLS}}$, and the corresponding model-assisted version of the RI estimator discussed in the supplementary material, $\widehat{\beta}^{w}_{\mathrm{RI}}$. For this latter estimator, we use all available covariates to adjust the standard estimators, that is, $\boldsymbol{W}$ is the entire vector of covariates.

Omnibus Test for Systematic Treatment Effect Variation. We begin by using these estimators for an omnibus test of whether any treatment effect variation is explained by the full set of covariates. The p-values for the unadjusted $\widehat{\beta}_{\mathrm{RI}}$ estimator and model-assisted $\widehat{\beta}^{w}_{\mathrm{RI}}$ are 0.39 and 0.25, respectively, which do not show any evidence of treatment effect variation. The OLS estimator, however, shows much stronger evidence with $p = 0.005$.

Importantly, all three estimators are based on the same underlying assumptions: the randomization itself justifies all three p-values. And while we expect the unadjusted $\widehat{\beta}_{\mathrm{RI}}$ to have



(a) Overall R^2_τ (b) R^2_τ separately by covariate

Figure 3. Treatment effect R^2_τ, with sensitivity parameter, $\rho \in [0, 1]$.

the lowest power, it is instructive that the p-value for $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is substantially smaller than the p-value for the covariate-adjusted $\widehat{\boldsymbol{\beta}}^{w}_{\text{RI}}$. As we discuss in Section 3.2, $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ can account for covariate imbalance across experimental arms by estimating the $\boldsymbol{S}_{xx}$ matrix separately for the treatment and control groups. By contrast, $\widehat{\boldsymbol{\beta}}_{\text{RI}}$ does not address imbalance in $\boldsymbol{X}$ and instead attempts to residualize out the Y to get a more precise estimate of the relationship of the $\boldsymbol{X}$ to Y for each treatment arm. Based on the discrepancy in p-values, adjusting for baseline imbalance is clearly important in this example.

Treatment Effect R^2_τ. Next, we examine how much of the variation could be explained by our covariates. Figure 3(a) shows values of the treatment effect R^2_τ using $\widehat{\boldsymbol{\beta}}^{w}_{\text{RI}}$ to estimate the systematic variation. Results are nearly identical using the other estimators. In the worst case of perfect negative dependence between potential outcomes (not shown), the treatment effect R^2_τ could be as low as 0.01. Assuming that this dependence is nonnegative, the treatment effect R^2_τ ranges from 0.03 to 0.76. While the estimate is clearly sensitive to the unidentifiable sensitivity parameter, the covariates explain a substantial proportion of treatment effect variation for values of ρ near 1.

We can also use this framework to assess the relative importance of each covariate in terms of explaining overall treatment effect variation. To do this, we use the model-assisted RI estimator, $\widehat{\boldsymbol{\beta}}^{w}_{\text{RI}}$, adjusting for all covariates (i.e., $\dim(\boldsymbol{W}) = 17$) but restricting systematic treatment effect variation to one covariate at a time. Note that we consider factors (e.g., race) as a group. Figure 3(b) shows the resulting estimates for the upper bound of R^2_τ, with lower bound estimates all below 0.01. Having a mother who is a recent immigrant and dual language learner status (which are highly correlated in practice) could each explain a substantial proportion of treatment effect variation, consistent with previous results from Bloom and Weiland (2014) and Bitler, Hoynes, and Domina (2014). This is not true for other covariates, like mother's education level.

Negative Correlation Between Treatment Effect and Control Potential Outcomes. Finally, we test whether the individual-level idiosyncratic treatment effects, $\{\varepsilon_i\}_{i=1}^n$, are negatively correlated with the control potential outcomes, $\{Y_i(0)\}_{i=1}^n$, extending results from Raudenbush and Bloom (2015). As outlined in the supplementary material, we do so by testing whether the variance of $\{Y_i^{\text{obs}} - \boldsymbol{X}_i^\intercal \widehat{\boldsymbol{\beta}}^{w}_{\text{RI}} : T_i = 1\}$ is smaller than the variance of $\{Y_i^{\text{obs}} : T_i = 0\}$. This yields a p-value of 0.02, which suggests that the unexplained treatment effect is indeed larger for smaller values of the control potential outcomes. This result is consistent with findings from Bitler, Hoynes, and Domina (2014) who use a quantile treatment effect approach.

7.2. Incorporating Noncompliance

As with many social experiments, there is substantial noncompliance with random assignment in HSIS. In the analysis sample we consider here, the estimated proportion of compliance types is $\widehat{\pi}_c = 0.69$ for Compliers, $\widehat{\pi}_a = 0.13$ for Always Takers, and $\widehat{\pi}_n = 0.18$ for Never Takers. Given the exclusion restrictions for Always Takers and Never Takers, the treatment effect is therefore zero (by assumption) for over 30% of the sample, suggesting that noncompliance will be an important component of treatment effect variation.

In the setting with noncompliance, we focus on two estimators for systematic treatment effect variation among Compliers: the randomization-based estimator, $\widehat{\boldsymbol{\beta}}_{c,\text{RI}}$, and the two-stage least-squares estimator, $\widehat{\boldsymbol{\beta}}_{\text{TSLS}}$. We first use these estimators to construct omnibus tests for systematic treatment effect variation among Compliers. Tests using both estimators show strong evidence for such variation, with p-value 0.02 using $\widehat{\boldsymbol{\beta}}_{c,\text{RI}}$ and p-value 0.01 using $\widehat{\boldsymbol{\beta}}_{\text{TSLS}}$.

Finally, we turn to decomposing the overall treatment effect. As in the ITT case, we assume that the potential outcomes have a nonnegative correlation. Figure 3(a) shows the treatment effect R^2 among Compliers, which ranges from $R^2_{\tau,c} = 0.05$ to $R^2_{\tau,c} = 0.68$. Next, we can calculate treatment effect variation due to noncompliance, $R^2_{\tau,U}$. In the case of HSIS, this is relatively small—between 0.01 and 0.16—in part because the overall treatment effect is fairly small. Therefore, the overall treatment effect decomposition due to both covariates and noncompliance, $R^2_{\tau,U\boldsymbol{X}}$, is quite close to $R^2_{\tau,c}$, as shown in Figure 3(a). Taken together, these estimates suggest that there is indeed important treatment effect variation that is neither captured by pretreatment covariates nor by noncompliance, consistent with previous results in Ding, Feller, and Miratrix (2016).

8. Conclusion

In this article, we propose a broad, flexible framework for assessing and decomposing treatment effect variation in randomized experiments with and without noncompliance. In general, we believe this is a natural setup for researchers to formulate and investigate a broad range of questions about impact heterogeneity (e.g., Heckman, Smith, and Clements 1997). Applications include assessing underlying causal mechanisms and targeting treatments based on individual-level characteristics. Understanding such variation is also important for the design of experiments. Djebbari and Smith (2008), for example, argued that characterizing the size of the idiosyncratic treatment effect is useful for determining the value of additional data collection.

We briefly note several directions for future work. First, our primary purpose was to propose a framework for analysis rooted in and justified by the randomization itself. As a result, we focused on the core properties of several relatively simple versions of linear regression and TSLS. We did not, however, fully explore their practical and finite-sample properties. For example, in future work, we hope to determine the settings in which model assistance will most improve estimation and assess the increased power of the OLS approach versus the unbiased RI approach. We are also investigating how to connect model-assisted and OLS approaches to take advantage of both methods of precision gain. Similarly, there is still much potential improvement in determining ways of characterizing the degree of heterogeneity, such as with an effect size for the systematic variation.

Second, a natural extension is to use more complex methods to estimate systematic treatment effects, such as via hierarchical models (Feller and Gelman 2015) or via machine-learning methods (Wager and Athey 2017), extending the results for the omnibus test and treatment effect R^2_τ accordingly. While the guarantees from randomization are clearly weaker in such

settings, researchers can assess these tradeoffs themselves. For example, hierarchical modeling would be especially useful in the Head Start Impact Study due to the multi-site design (Bloom and Weiland 2014).

Third, a question of increasing practical importance is the generalizability of experimental results to a given target population (Stuart et al. 2011). We believe that the treatment effect R^2_τ is a critical measure for assessing the credibility of these generalizations. In short, if there is substantial idiosyncratic treatment effect variation, that is, R^2_τ is small, then researchers should be wary of using observed covariates to extrapolate treatment effects.

Finally, a question is how to extend this treatment effect variation framework to nonrandomized settings. While the results would necessarily rest on much stronger assumptions, many settings already use an as-if-randomized framework, such as in observational studies (Rosenbaum 2002; Imbens and Rubin 2015). Under this approach, extensions should be natural.

Acknowledgments

The authors thank Alberto Abadie, Donald Rubin, participants at the Applied Statistics Seminar at the Harvard Institute of Quantitative Social Science, and colleagues at University of California, Berkeley and Harvard University for helpful comments. The authors also thank their reviewers who helped them sharpen their mathematical presentations, in particular the asymptotic arguments.

Funding

The authors 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). Peng Ding also gratefully acknowledges financial support from the National Science Foundation (DMS grant #1713152).

References

Abadie, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263. [304,310]

Administration for Children and Families (2015), "Head Start Program Facts, Fiscal Year 2014," available at *https://eclkc.ohs.acf.hhs.gov/hslc/data/factsheets/docs/hs-program-fact-sheet-2014.pdf* [313]

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455. [304,310]

Angrist, J. D., Pathak, P. A., and Walters, C. R. (2013), "Explaining Charter School Effectiveness," *American Economic Journal: Applied Economics*, 5, 1–27. [305]

Angrist, J. D., and Pischke, J. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press. [307,310,311]

Aronow, P. M., Green, D. P., and Lee, D. K. (2014), "Sharp Bounds on the Variance in Randomized Experiments," *The Annals of Statistics*, 42, 850–871. [307,309]

Athey, S., and Imbens, G. (2016), "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360. [304]

Berrington de González, A., and Cox, D. R. (2007), "Interpretation of Interaction: A Review," *The Annals of Applied Statistics*, 1, 371–385. [307]

Bitler, M., Hoynes, H., and Domina, T. (2014), "Experimental Evidence on Distributional Effects of Head Start," NBER Working Paper 20434. [313,315]

Blinder, A. S. (1973), "Wage Discrimination: Reduced form and Structural Estimates," *Journal of Human resources*, 8, 436–455. [305]

Bloom, H. S., and Weiland, C. (2015), "Qualifying Variation in Head Start Effects on Young Children's Cognitive and Socio-Emotional Skills Using Data from the National Head Start Impact Study," SSRN Working Paper 2594430. [313,315]

Bound, J., Jaeger, D. A., and Baker, R. M. (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. [310]

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley. [5]

Cox, D. R. (1984), "Interaction" (with discussion), *International Statistical Review*, 52, 1–24. [304,307]

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008), "Nonparametric Tests for Treatment Effect Heterogeneity," *Review of Economics and Statistics*, 90, 389–405. [304,307,308]

Currie, J., and Thomas, D. (1995), "Does Head Start Make a Difference?" *American Economic Review*, 85, 341–364. [313]

Ding, P. (2017), "A Paradox from Randomization-Based Causal Inference" (with discussion), *Statistical Science*, 32, 331–335. [307]

Ding, P., Feller, A., and Miratrix, L. W. (2016), "Randomization Inference for Treatment Effect Variation," *Journal of the Royal Statistical Society*, Series B, 78, 655–671. [308,313,314,315]

Djebbari, H., and Smith, J. (2008), "Heterogeneous Impacts in PROGRESA," *Journal of Econometrics*, 145, 64–80. [304,306,309,315]

Feller, A., and Gelman, A. (2015), "Hierarchical Models for Causal Effects," in *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, eds. R. Scott and S. Kosslyn, New York: Wiley. [315]

Feller, A., Grindal, T., Miratrix, L., and Page, L. C. (2016), "Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care Type," *The Annals of Applied Statistics*, 10, 1245–1285. [313]

Fisher, R. A. (1935), *The Design of Experiments* (1st ed.), Edinburgh: Oliver & Boyd. [304]

Fogarty, C. B. (forthcoming), "Regression Assisted Inference for the Average Treatment Effect in Paired Experiments," *Biometrika*. [309]

Frangakis, C. E., and Rubin, D. B. (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29. [310]

Fréchet, M. (1951), "Sur Les Tableaux de Corrélation dont les Marges son Données," *Annals Universite de Lyon, Section A, Series 3*, 14, 53–77. [308]

Green, D. P., and Kern, H. L. (2012), "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees," *The Public Opinion Quarterly*, 76, 491–511. [304]

Hájek, J. (1960), "Limiting Distributions in Simple Random Sampling from a Finite Population," *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5, 361–374. [307]

Heckman, J. J., Smith, J., and Clements, N. (1997), "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64, 487–535. [304,306,309,315]

Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240. [304]

Hoeffding, W. (1941), "Masstabinvariante Korrelationsmasse Für Diskontinuierliche Verteilungen," *Arkiv fr matematischen Wirtschaften und Sozialforschung*, 7, 49–70. [308]

Huang, Y., Gilbert, P. B., and Janes, H. (2012), "Assessing Treatment-Selection Markers Using a Potential Outcomes Framework," *Biometrics*, 68, 687–696. [304]

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), pp. 221–233. Berkeley: University of California Press. [307]

——— (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799–821. [307]

Imai, K., and Ratkovic, M. (2013), "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation," *The Annals of Applied Statistics*, 7, 443–470. [304]

Imbens, G. (2014), "Instrumental Variables: An Econometrician's Perspective" (with discussion), *Statistical Science*, 29, 323–358. [304,311]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, and in the Social and Biomedical Sciences*, New York: Cambridge University Press. [304,305,310,316]

Kempthorne, O. (1952), *The Design and Analysis of Experiments*, New York: Wiley. [304]

Lehmann, E. L. (1998), *Elements of Large-Sample Theory*, New York: Springer. [307]

Li, X., and Ding, P. (2017), "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference," *Journal of the American Statistical Association*, 112, 1759–1769. [305,307]

Lin, W. (2013), "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique," *The Annals of Applied Statistics*, 7, 295–318. [304,306,307]

Matsouaka, R. A., Li, J., and Cai, T. (2014), "Evaluating Marker-Guided Treatment Selection Strategies," *Biometrics*, 70, 489–499. [304]

Middleton, J. A., and Aronow, P. M. (2015), "Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments," *Statistics, Politics and Policy*, 6, 39–75. [306,307]

Nelsen, R. B. (2007), *An Introduction to Copulas* (2nd ed.), New York: Springer. [308,309]

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statistical Science*, 5, 465–472. [304,305,306,309]

Oaxaca, R. (1973), "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14, 693–709. [305]

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., Jenkins, F., Fletcher, P., Quinn, L., Friedman, J., Ciarico, J., Rohacek, M., Adams, M., and Spier, E. (2010), "Head Start Impact Study: Final Report," Technical Report, Department of Health and Human Services, Administration for Children and Families, Washington DC. [305,313,314]

Raudenbush, S. W., and Bloom, H. S. (2015), "Learning About and from a Distribution of Program Impacts Using Multisite Trials," *American Journal of Evaluation*, 36, 475–499. [304,309,315]

Rosenbaum, P. R. (1999), "Reduced Sensitivity to Hidden Bias at Upper Quantiles in Observational Studies with Dilated Treatment Effects," *Biometrics*, 55, 560–564. [304]

——— (2002), *Observational Studies* (2nd ed.), New York: Springer. [304,305,316]

——— (2007), "Confidence Intervals for Uncommon but Dramatic Responses to Treatment," *Biometrics*, 63, 1164–1171. [304]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [305]

——— (1980), "Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test" by D. Basu," *Journal of the American Statistical Association*, 75, 591–593. [305]

Särndal, C.-E., Swensson, B., and Wretman, J. (2003), *Model-Assisted Survey Sampling*, New York: Springer. [308]

Staiger, D. O., and Stock, J. H. (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586. [310]

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011), "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials," *Journal of the Royal Statistical Society*, Series A, 174, 369–386. [316]

Wager, S., and Athey, S. (2017), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, doi:10.1080/01621459.2017.1319839. [304,315]

Walters, C. R. (2015), "Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start," *American Economic Journal: Applied Economics*, 7, 76–102. [313]

Westinghouse Learning Corporation (1969), *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development, Volume 1: Report to the Office of Economic Opportunity*, Athens, OH: Westinghouse Learning Corporation and Ohio University. [313]

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, *48*, 817–838. [307]

Zhao, A., Ding, P., Mukerjee, R., and Dasgupta, T. (in press), "Randomization-Based Causal Inference from Split-Plot Designs," *Annals of Statistics*. [306]

Supplementary Material for "Decomposing Treatment Effect Variation" by Peng Ding, Avi Feller and Luke Miratrix

Appendix A gives all the proofs and Appendix B provides the additional commentary mentioned in the main text. The finite population central limit theorem (FPCLT) we use for our asymptotic proofs is Theorem 5 of Li and Ding (2017), which requires some mild moment conditions on the covariates and potential outcomes, as outlined in the main text.

Appendix A Lemmas and Proofs

Before we prove Theorem 1, we provide a few lemmas to ease the notational burden and amount of algebra of subsequent calculations. These lemmas allow us to derive expressions for our estimators in terms of matrix algebra rather than the summation-style approach typically seen for Neyman-style derivations in the literature.

To begin, let $\mathbf{1}_n = (1, \ldots, 1)^\intercal$ and $\mathbf{0}_n = (0, \ldots, 0)^\intercal$ be column vectors of length n, and $\boldsymbol{I}_n$ be the $n \times n$ identity matrix. Then $\boldsymbol{S}_n = \boldsymbol{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\intercal$ is the projection matrix orthogonal to $\mathbf{1}_n$ with $\boldsymbol{S}_n\mathbf{1}_n = \mathbf{0}_n$. Under this formulation, the covariance matrix of the treatment assignment vector is a scaled projection matrix orthogonal to $\mathbf{1}_n$, as shown in the following lemma.

Lemma A.1. The treatment assignment vector $\boldsymbol{T}$ of a completely randomized experiment has

$$E(\boldsymbol{T}) = \frac{n_1}{n}1_n, \quad \text{cov}(\boldsymbol{T}) = \frac{n_1 n_0}{n(n-1)}\boldsymbol{S}_n.$$

Proof of Lemma A.1. The conclusions follow from

$$E(T_i) = \frac{n_1}{n}, \quad \text{var}(T_i) = \frac{n_1 n_0}{n^2}, \quad \text{cov}(T_i, T_j) = -\frac{n_1 n_0}{n^2(n-1)}, \quad (i \neq j).$$

□

The projection matrix $\boldsymbol{S}_n$ acts as a covariance operator as illustrated by the following lemma.

Lemma A.2. Let $\boldsymbol{U}_i, \boldsymbol{V}_i \in \mathbb{R}^K$ be column vectors of length K. Define $\mathcal{U} = [\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_n]$ and $\mathcal{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_n] \in \mathbb{R}^{K \times n}$ as two matrices of dimension $K \times n$. If $\bar{\boldsymbol{U}} = n^{-1}\sum_{i=1}^n \boldsymbol{U}_i = n^{-1}\mathcal{U}\mathbf{1}$ and $\bar{\boldsymbol{V}} = n^{-1}\sum_{i=1}^n \boldsymbol{V}_i = n^{-1}\mathcal{V}\mathbf{1}$, then

$$\mathcal{U}\boldsymbol{S}_n\mathcal{V}^\intercal = \sum_{i=1}^n (\boldsymbol{U}_i - \bar{\boldsymbol{U}})(\boldsymbol{V}_i - \bar{\boldsymbol{V}})^\intercal. \tag{A.1}$$

In particular, when $\boldsymbol{U}_i = \boldsymbol{V}_i$,

$$\mathcal{V}\boldsymbol{S}_n\mathcal{V}^{\intercal} = \sum_{i=1}^{n}(\boldsymbol{V}_i - \bar{\boldsymbol{V}})(\boldsymbol{V}_i - \bar{\boldsymbol{V}})^{\intercal} = (n-1)\mathcal{S}(\boldsymbol{V}).$$

Proof of Lemma A.2. The left hand side of (A.1) is equal to

$$\mathcal{U}\boldsymbol{S}_n\mathcal{V}^{\intercal} = \mathcal{U}\mathcal{V}^{\intercal} - n^{-1}\left(\mathcal{U}\mathbf{1}_n\right)\left(\mathcal{V}\mathbf{1}_n\right)^{\intercal} = \sum_{i=1}^{n}\boldsymbol{U}_i\boldsymbol{V}_i^{\intercal} - n^{-1}(n\bar{\boldsymbol{U}})(n\bar{\boldsymbol{V}})^{\intercal} = \sum_{i=1}^{n}\boldsymbol{U}_i\boldsymbol{V}_i^{\intercal} - n\bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^{\intercal},$$

which is the same as the right hand side of (A.1). □

Theorem 1: A generalized, vector-outcome version of Neyman. To prove the generalized Neyman result, we bundle our vector potential outcomes into matrices and use the above lemmas to obtain their covariance matrix. The theorem is exact, no asymptotics. Using the FPCLT to show that the estimator has an approximately Normal distribution, allowing for classic testing and inference, is a separate, subsequent step.

Proof of Theorem 1. Define $\mathcal{V}_1 = [\boldsymbol{V}_1(1), \ldots, \boldsymbol{V}_n(1)]$ and $\mathcal{V}_0 = [\boldsymbol{V}_1(0), \ldots, \boldsymbol{V}_n(0)]$ as the matrices of the potential outcomes. Then the Neymanian simple difference in means estimator has the following representation:

$$\begin{aligned}
\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}} &= \bar{\boldsymbol{V}}_1^{\text{obs}} - \bar{\boldsymbol{V}}_0^{\text{obs}} \\
&= \frac{1}{n_1}\sum_{i=1}^{n} T_i\boldsymbol{V}_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)\boldsymbol{V}_i(0) \\
&= \frac{1}{n_1}\mathcal{V}_1\boldsymbol{T} - \frac{1}{n_0}\mathcal{V}_0\left(\mathbf{1} - \boldsymbol{T}\right) \\
&= \left(\frac{\mathcal{V}_1}{n_1} + \frac{\mathcal{V}_0}{n_0}\right)\boldsymbol{T} - \frac{1}{n_0}\mathcal{V}_0\mathbf{1}.
\end{aligned}$$

Now the unbiasedness of $\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}}$ follows from the linearity of the expectation and Lemma A.1. For the covariance, note the second term in the above is constant, and so is not involved. Applying Lemmas A.1 and A.2, we can obtain the covariance matrix of $\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}}$:

$$\begin{aligned}
\text{cov}(\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}}) &= \left(\frac{\mathcal{V}_1}{n_1} + \frac{\mathcal{V}_0}{n_0}\right)\text{cov}(\boldsymbol{T})\left(\frac{\mathcal{V}_1}{n_1} + \frac{\mathcal{V}_0}{n_0}\right)^{\intercal} \\
&= \frac{n_1 n_0}{n(n-1)}\left(\frac{\mathcal{V}_1}{n_1} + \frac{\mathcal{V}_0}{n_0}\right)\boldsymbol{S}_n\left(\frac{\mathcal{V}_1}{n_1} + \frac{\mathcal{V}_0}{n_0}\right)^{\intercal} \\
&= \frac{n_1 n_0}{n(n-1)}\left(\frac{1}{n_1^2}\mathcal{V}_1\boldsymbol{S}_n\mathcal{V}_1^{\intercal} + \frac{1}{n_0^2}\mathcal{V}_0\boldsymbol{S}_n\mathcal{V}_0^{\intercal} + \frac{1}{n_1 n_0}\mathcal{V}_0\boldsymbol{S}_n\mathcal{V}_1^{\intercal} + \frac{1}{n_1 n_0}\mathcal{V}_1\boldsymbol{S}_n\mathcal{V}_0^{\intercal}\right) \\
&= \frac{n_0}{n n_1}\mathcal{S}\{\boldsymbol{V}(1)\} + \frac{n_1}{n n_0}\mathcal{S}\{\boldsymbol{V}(0)\} + \frac{1}{n(n-1)}(\mathcal{V}_0\boldsymbol{S}_n\mathcal{V}_1^{\intercal} + \mathcal{V}_1\boldsymbol{S}_n\mathcal{V}_0^{\intercal}).
\end{aligned}$$

To simplify the third term, we use the fact $\boldsymbol{a}\boldsymbol{b}^{\intercal} + \boldsymbol{b}\boldsymbol{a}^{\intercal} = \boldsymbol{a}\boldsymbol{a}^{\intercal} + \boldsymbol{b}\boldsymbol{b}^{\intercal} - (\boldsymbol{a}-\boldsymbol{b})(\boldsymbol{a}-\boldsymbol{b})^{\intercal}$ for two column vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, we have

$$\begin{aligned} & \{\boldsymbol{V}_i(1) - \bar{\boldsymbol{V}}(1)\}\{\boldsymbol{V}_i(0) - \bar{\boldsymbol{V}}(0)\}^{\intercal} + \{\boldsymbol{V}_i(0) - \bar{\boldsymbol{V}}(0)\}\{\boldsymbol{V}_i(1) - \bar{\boldsymbol{V}}(1)\}^{\intercal} \\ = \; & \{\boldsymbol{V}_i(1) - \bar{\boldsymbol{V}}(1)\}\{\boldsymbol{V}_i(1) - \bar{\boldsymbol{V}}(1)\}^{\intercal} + \{\boldsymbol{V}_i(1) - \bar{\boldsymbol{V}}(1)\}\{\boldsymbol{V}_i(1) - \bar{\boldsymbol{V}}(1)\}^{\intercal} \\ & -\{\boldsymbol{V}_i(1) - \boldsymbol{V}_i(0) - \bar{\boldsymbol{V}}(1) + \bar{\boldsymbol{V}}(0)\}\{\boldsymbol{V}_i(1) - \boldsymbol{V}_i(0) - \bar{\boldsymbol{V}}(1) + \bar{\boldsymbol{V}}(0)\}^{\intercal}. \end{aligned}$$

Summing over $i = 1, \ldots, n$ and applying Lemma A.2, we have

$$\frac{\mathcal{V}_0 \boldsymbol{S}_n \mathcal{V}_1^{\intercal}}{n-1} + \frac{\mathcal{V}_1 \boldsymbol{S}_n \mathcal{V}_0^{\intercal}}{n-1} = \mathcal{S}\{\boldsymbol{V}(1)\} + \mathcal{S}\{\boldsymbol{V}(0)\} - \mathcal{S}\{\boldsymbol{V}(1) - \boldsymbol{V}(0)\}.$$

Therefore, the covariance of $\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}}$ can be simplified as:

$$\begin{aligned} \operatorname{cov}(\widehat{\boldsymbol{\tau}}_{\boldsymbol{V}}) &= \frac{n_0}{n n_1}\mathcal{S}\{\boldsymbol{V}(1)\} + \frac{n_1}{n n_0}\mathcal{S}\{\boldsymbol{V}(0)\} + \frac{1}{n}\left[\mathcal{S}\{\boldsymbol{V}(1)\} + \mathcal{S}\{\boldsymbol{V}(0)\} - \mathcal{S}\{\boldsymbol{V}(1) - \boldsymbol{V}(0)\}\right] \\ &= \frac{\mathcal{S}\{\boldsymbol{V}(1)\}}{n_1} + \frac{\mathcal{S}\{\boldsymbol{V}(0)\}}{n_0} - \frac{\mathcal{S}\{\boldsymbol{V}(1) - \boldsymbol{V}(0)\}}{n}. \end{aligned}$$

□

Theorem 2: Behavior of $\widehat{\boldsymbol{\beta}}_{\mathbf{RI}}$. To show properties of $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ we express the systematic variation as a vector of new potential outcomes of the original outcome scaled by the different covariates of interest. This allows for immediate use of Theorem 1.

Proof of Theorem 2. Because $\widehat{\boldsymbol{S}}_{xt}$ is the sample mean for $\{\boldsymbol{X}_i Y_i^{\mathrm{obs}} : T_i = t, i = 1, \ldots, n\} = \{\boldsymbol{X}_i Y_i(t) : T_i = t, i = 1, \ldots, n\}$, it is unbiased for the population mean $\boldsymbol{S}_{xt}$. Thus, the estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ is also unbiased for $\boldsymbol{\beta}$ as $\boldsymbol{S}_{xx}^{-1}$ is fixed and the expectation is linear. Its sampling covariance over all possible randomizations is

$$\operatorname{cov}(\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}) = \boldsymbol{S}_{xx}^{-1} \operatorname{cov}(\widehat{\boldsymbol{S}}_{x1} - \widehat{\boldsymbol{S}}_{x0}) \boldsymbol{S}_{xx}^{-1}.$$

Therefore, we need only to obtain the covariance of

$$\widehat{\boldsymbol{S}}_{x1} - \widehat{\boldsymbol{S}}_{x0} = \frac{1}{n_1}\sum_{i=1}^{n} T_i \boldsymbol{X}_i Y_i^{\mathrm{obs}} - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_i)\boldsymbol{X}_i Y_i^{\mathrm{obs}},$$

which is the difference between the sample means of $\{\boldsymbol{X}_i Y_i(1) : i = 1, \ldots, n\}$ and $\{\boldsymbol{X}_i Y_i(0) : i = 1, \ldots, N\}$ under treatment and control. Viewing $\boldsymbol{X}_i Y_i^{\mathrm{obs}}$ as a vector outcome in a completely randomized experiment, we can apply Theorem 1 to obtain

$$\operatorname{cov}(\widehat{\boldsymbol{S}}_{x1} - \widehat{\boldsymbol{S}}_{x0}) = \frac{\mathcal{S}\{\boldsymbol{X}Y(1)\}}{n_1} + \frac{\mathcal{S}\{\boldsymbol{X}Y(0)\}}{n_0} - \frac{\mathcal{S}(\boldsymbol{X}\tau)}{n},$$

which completes the proof. □

Theorem 3: Behavior of $\widehat{\boldsymbol{\beta}}_{\mathbf{OLS}}$. We first use the well-known fact that the estimate from a OLS model with treatment fully interacted with covariates is equivalent to separate regressions of outcome onto covariates for the control and treatment groups. This means we can obtain $\widehat{\boldsymbol{\gamma}}_{\mathrm{OLS}}$ by running a regression of Y^{obs} onto $\boldsymbol{X}$ using the control group data, and $\widehat{(\boldsymbol{\gamma}+\boldsymbol{\beta})}_{\mathrm{OLS}}$ by running regression of Y^{obs} onto $\boldsymbol{X}$ using the treatment group data, giving estimated coefficients of

$$\widehat{\boldsymbol{\gamma}}_{\mathrm{OLS}} = \widehat{\boldsymbol{S}}_{xx,0}^{-1}\widehat{\boldsymbol{S}}_{x0}$$

and

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \widehat{\boldsymbol{S}}_{xx,1}^{-1}\widehat{\boldsymbol{S}}_{x1} - \widehat{\boldsymbol{S}}_{xx,0}^{-1}\widehat{\boldsymbol{S}}_{x0}.$$

As a quick heuristic argument for this, consider that the maximization problem for the interacted model will separate into two components, one for each group. Then re-parameterize to get the above.

We now prove the properties of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. Here we have to use asymptotics for the entire theorem, unlike the case of $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$, where the mean and covariance are exact and the asymptotics are only needed for the asymptotic normality of the estimator.

Proof of Theorem 3. First expand the difference of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta} = \widehat{\boldsymbol{S}}_{xx,1}^{-1}(\widehat{\boldsymbol{S}}_{x1} - \widehat{\boldsymbol{S}}_{xx,1}\boldsymbol{\gamma}_1) - \widehat{\boldsymbol{S}}_{xx,0}^{-1}(\widehat{\boldsymbol{S}}_{x0} - \widehat{\boldsymbol{S}}_{xx,0}\boldsymbol{\gamma}_0),$$

This will be close to the related quantity of

$$\Delta = \boldsymbol{S}_{xx}^{-1}(\widehat{\boldsymbol{S}}_{x1} - \widehat{\boldsymbol{S}}_{xx,1}\gamma_1) - \boldsymbol{S}_{xx}^{-1}(\widehat{\boldsymbol{S}}_{x0} - \widehat{\boldsymbol{S}}_{xx,0}\boldsymbol{\gamma}_0). \tag{A.2}$$

For the above to make sense and hold, we here need our asymptotic framework. In particular, we need the associated moment conditions described in the main text. We next observe that the difference between $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}$ and Δ is of higher order, because

$$\begin{aligned}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \beta) - \Delta &= (\widehat{\boldsymbol{S}}_{xx,1}^{-1} - \boldsymbol{S}_{xx}^{-1})(\widehat{\boldsymbol{S}}_{x1} - \widehat{\boldsymbol{S}}_{xx,1}\boldsymbol{\gamma}_1) - (\widehat{\boldsymbol{S}}_{xx,0}^{-1} - \boldsymbol{S}_{xx}^{-1})(\widehat{\boldsymbol{S}}_{x0} - \widehat{\boldsymbol{S}}_{xx,0}\boldsymbol{\gamma}_0) \quad \text{(A.3)}\\ &= O_P(n^{-1/2})O_P(n^{-1/2}) - O_P(n^{-1/2})O_P(n^{-1/2}) = O_P(n^{-1}), \quad \text{(A.4)}\end{aligned}$$

following from the FPCLT for the four terms in (A.3). This is an argument commonly used in the survey sampling literature for ratio estimators (Cochran, 1977).

We next focus on the asymptotic distribution of Δ, because the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta}$ will be the same. Further simplify (A.2) as

$$\Delta = \boldsymbol{S}_{xx}^{-1} \left[\frac{1}{n_1} \sum_{i=1}^{n} T_i \boldsymbol{X}_i e_i(1) - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) \boldsymbol{X}_i e_i(0) \right], \tag{A.5}$$

where $e_i(1) = Y_i(1) - \boldsymbol{X}_i^{\intercal} \boldsymbol{\gamma}_1$ and $e_i(0) = Y_i(0) - \boldsymbol{X}_i^{\intercal} \boldsymbol{\gamma}_0$ are the residual potential outcomes. (To see the above, note, for example, that both $\widehat{\boldsymbol{S}}_{x1}$ and $\widehat{\boldsymbol{S}}_{xx,1}$ are sums over the treatment units, and we can factor out an $\boldsymbol{X}_i$ to get $\boldsymbol{X}_i$ times the difference in the Y_i and predicted Y_i.)

Applying Theorem 1 to the vector outcome $\boldsymbol{X}e$, we obtain the covariance matrix of Δ, to which $\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \beta$ converges to due to (A.4). The asymptotic normality follows from the representation (A.5) and the FPCLT. □

Theorem 4: Bounds for R_τ^2. To prove Theorem 4, we need to invoke the following Fréchet–Hoeffding inequality (Hoeffding, 1941; Fréchet, 1951; Heckman et al., 1997; Aronow et al., 2014).

Lemma A.3. If we know only the marginal distributions of two random variables $X \sim F_X(x)$ and $Y \sim F_Y(y)$, then $E(XY)$ can be sharply bounded by

$$\int_0^1 F_X^{-1}(u) F_Y^{-1}(1-u) \mathrm{d}u \leq E(XY) \leq \int_0^1 F_X^{-1}(u) F_Y^{-1}(u) \mathrm{d}u.$$

Lemma A.3 immediately implies the following bound for $\text{var}(X - Y)$ if $E(X - Y) = 0$.

Lemma A.4. If we know only the marginal distributions $X \sim F_X(x), Y \sim F_Y(y)$ and $E(X - Y) = 0$, then $\text{var}(X - Y)$ can be sharply bounded by

$$\int_0^1 \{F_X^{-1}(u) - F_Y^{-1}(u)\}^2 \mathrm{d}u \leq \text{var}(X - Y) \leq \int_0^1 \{F_X^{-1}(u) - F_Y^{-1}(1-u)\}^2 \mathrm{d}u$$

Proof of Lemma A.4. The variance $\text{var}(X - Y)$ can be decomposed as

$$\text{var}(X - Y) = E(X - Y)^2 = E(X^2) + E(Y^2) - 2E(XY),$$

which depends on the following three terms:

$$\begin{aligned} E(X^2) = & \quad \int x^2 \mathrm{d}F_X(x) & = \int_0^1 \{F_X^{-1}(u)]\}^2 \mathrm{d}u, \\ E(Y^2) = & \quad \textstyle\int_0^1 \{F_Y^{-1}(u)\}^2 \mathrm{d}u & = \displaystyle\int_0^1 \{F_Y^{-1}(1-u)\}^2 \mathrm{d}u, \\ \int_0^1 F_X^{-1}(u) F_Y^{-1}(1-u) \mathrm{d}u \leq & \quad E(XY) & \leq \int_0^1 F_X^{-1}(u) F_Y^{-1}(u) \mathrm{d}u. \end{aligned}$$

Plug the above expressions into the variance of $X - Y$ to obtain the desired bounds. □

Applying Lemma A.4, we can easily prove Theorem 4.

Proof of Theorem 4. Because $S_{\tau\tau} = S_{\delta\delta} + S_{\varepsilon\varepsilon}$, we need only to bound $S_{\varepsilon\varepsilon}$, which is the finite population variance of $\varepsilon_i = \{Y_i(1) - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\gamma}_1\} - \{Y_i(0) - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\gamma}_0\} = e_i(1) - e_i(0)$. We can identify the marginal distributions of $\{e_i(1) : i = 1, \ldots, n\}$ and $\{e_i(0) : i = 1, \ldots, n\}$, and also know that $n^{-1}\sum_{i=1}^n \varepsilon_i = 0$. Therefore, the bounds in Lemma A.4 imply the bounds in Theorem 4. □

Theorem 5: Sensitivity analysis.

Proof of Theorem 5. The joint distribution of (U_1, U_0) is

$$\begin{aligned} C(u_1, u_0) &= P(U_1 \le u_1, U_0 \le u_0) \\ &= \rho P(U_0 \le u_1, U_0 \le u_0) + (1-\rho)P(V_0 \le u_1, U_0 \le u_0) \\ &= \rho \min(u_1, u_0) + (1-\rho)u_1 u_0. \end{aligned}$$

Therefore, the distribution function $C(u_1, u_0)$ is a weighted average of $\min(u_1, u_0) = C_R(u_1, u_0)$ and $u_1 u_0 = C_I(u_1, u_0)$, i.e., the joint distributions when $U_1 = U_0$ and $U_1 \perp\!\!\!\perp U_0$, respectively.

According to Nelsen (2007, Theorem 5.1.6), Spearman's rank correlation coefficient between $e(1)$ and $e(0)$ is

$$\begin{aligned} 12\int_0^1\int_0^1 \{C(u_1, u_0) - u_1 u_0\}\mathrm{d}u_1\mathrm{d}u_0 &= 12\rho\int_0^1\int_0^1 \{\min(u_1, u_0) - u_1 u_0\}\mathrm{d}u_1\mathrm{d}u_0 \\ &= 12\rho\left(2\int_0^1 \mathrm{d}u_1\int_0^{u_1} u_0\mathrm{d}u_0 - \frac{1}{4}\right) \\ &= 12\rho(1/3 - 1/4) = \rho. \end{aligned}$$

To complete the proof of the theorem, we need only to show that the covariance between $e(1)$ and $e(0)$ is linear in ρ, which follows from

$$\begin{aligned} &\int_0^1\int_0^1 F_1^{-1}(u_1)F_0^{-1}(u_0)\mathrm{d}C(u_1, u_0) \\ =\ & \rho\int_0^1\int_0^1 F_1^{-1}(u_1)F_0^{-1}(u_0)\mathrm{d}C_R(u_1, u_0) + \rho\int_0^1\int_0^1 F_1^{-1}(u_1)F_0^{-1}(u_0)\mathrm{d}C_I(u_1, u_0) \\ =\ & \rho\int_0^1 F_1^{-1}(u)F_0^{-1}(u)\mathrm{d}u + (1-\rho)\int_0^1 F_1^{-1}(u)\mathrm{d}u\int_0^1 F_0^{-1}(u)\mathrm{d}u. \end{aligned}$$

□

Theorem 6: Extending to non-compliance. Theorem 6 shows how to estimate the outcome-to-covariate relationships of the Compliers by estimating different aggregate covariance relationships across all the strata for different observed groups and then taking differences. Due to the exclusion restriction for the Never Takers and Always Takers, this gives our desired relationships for the Compliers only.

First, a small bit of notation of, due to the exclusion restrictions for Never Takers and Always Takers, defining the population covariance between $\boldsymbol{X}$ and $Y(1) = Y(0)$ within stratum $U = a$ and $U = n$ as

$$\boldsymbol{S}_{x.,u} = \frac{1}{n_u}\sum_{i=1}^{n} I_{(U_i=u)}\boldsymbol{X}_i Y_i(1) = \frac{1}{n_u}\sum_{i=1}^{n} I_{(U_i=u)}\boldsymbol{X}_i Y_I(0), \quad (u = a, n).$$

Proof of Theorem 6. We first create an estimator for $\boldsymbol{S}_{xx,c}$. From the observed data with $(T_i, D_i) = (1, 1)$, we have

$$\begin{aligned} E\left\{\frac{1}{n_1}\sum_{i=1}^{n} T_i D_i \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}}\right\} &= E\left\{\frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=a)}\boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}} + \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=c)}\boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}}\right\} \\ &= \pi_a \boldsymbol{S}_{xx,a} + \pi_c \boldsymbol{S}_{xx,c}. \end{aligned} \tag{A.6}$$

Similar to (A.6), we have

$$E\left\{\frac{1}{n_1}\sum_{i=1}^{n} T_i(1 - D_i)\boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}}\right\} = \pi_n \boldsymbol{S}_{xx,n}, \tag{A.7}$$

$$E\left\{\frac{1}{n_0}\sum_{i=1}^{n} (1 - T_i) D_i \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}}\right\} = \pi_a \boldsymbol{S}_{xx,a}, \tag{A.8}$$

$$E\left\{\frac{1}{n_0}\sum_{i=1}^{n} (1 - T_i)(1 - D_i)\boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}}\right\} = \pi_n \boldsymbol{S}_{xx,n} + \pi_c \boldsymbol{S}_{xx,c}. \tag{A.9}$$

Subtracting the left sides of (A.8) from (A.6), or subtracting the left sides of (A.7) from (A.9), give unbiased estimators for $\pi_c \boldsymbol{S}_{xx,c}$.

Second, analogous to the $\boldsymbol{S}_{xx,c}$, we consider the sample covariances between $\boldsymbol{X}$ and Y^{obs} to obtain estimators for $\boldsymbol{S}_{x1,c}$ and $\boldsymbol{S}_{x0,c}$. From the observed data with $(T_i, D_i) = (1, 1)$, we have

$$\begin{aligned} E\left\{\frac{1}{n_1}\sum_{i=1}^{n} T_i D_i \boldsymbol{X}_i Y_i^{\mathrm{obs}}\right\} &= E\left\{\frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=a)}\boldsymbol{X}_i Y_i(1) + \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=c)}\boldsymbol{X}_i Y_i(1)\right\} \\ &= \pi_a \boldsymbol{S}_{x.,a} + \pi_c \boldsymbol{S}_{x1,c}. \end{aligned} \tag{A.10}$$

Similar to (A.10), we have

$$E\left\{\frac{1}{n_1}\sum_{i=1}^{n} T_i(1-D_i)\boldsymbol{X}_i Y_i^{\text{obs}}\right\} = \pi_n \boldsymbol{S}_{x.,n}, \tag{A.11}$$

$$E\left\{\frac{1}{n_0}\sum_{i=1}^{n} (1-T_i)D_i\boldsymbol{X}_i Y_i^{\text{obs}}\right\} = \pi_a \boldsymbol{S}_{x.,a}, \tag{A.12}$$

$$E\left\{\frac{1}{n_0}\sum_{i=1}^{n} (1-T_i)(1-D_i)\boldsymbol{X}_i Y_i^{\text{obs}}\right\} = \pi_n \boldsymbol{S}_{x.,n} + \pi_c \boldsymbol{S}_{x0,c}. \tag{A.13}$$

Subtracting (A.12) from (A.10), and subtracting (A.11) from (A.13), we obtain the results in (15). □

Corollary 3: Behavior of $\widehat{\boldsymbol{\beta}}_{c,\mathbf{RI}}$. Theorem 6 shows how to obtain unbiased estimates of the components of our estimator, which we can then plug in to obtain a consistent estimator of $\boldsymbol{\beta}_c$. We next show how this plug-in estimator behaves.

Proof of Corollary 3. First we write

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{c,\text{RI}} - \boldsymbol{\beta}_c &= (\widehat{\boldsymbol{S}}_{xx,11} - \widehat{\boldsymbol{S}}_{xx,01})^{-1}\{\widehat{\boldsymbol{S}}_{x1,11} - \widehat{\boldsymbol{S}}_{x0,01} - (\widehat{\boldsymbol{S}}_{xx,11} - \widehat{\boldsymbol{S}}_{xx,01})\boldsymbol{\gamma}_{1c}\} \\ &\quad -(\widehat{\boldsymbol{S}}_{xx,00} - \widehat{\boldsymbol{S}}_{xx,10})^{-1}\{\widehat{\boldsymbol{S}}_{x0,00} - \widehat{\boldsymbol{S}}_{x1,10} - (\widehat{\boldsymbol{S}}_{xx,00} - \widehat{\boldsymbol{S}}_{xx,10})\boldsymbol{\gamma}_{0c}\},\end{aligned}$$

second we introduce

$$\begin{aligned}\Delta_c &= (\pi_c \boldsymbol{S}_{xx,c})^{-1}\{\widehat{\boldsymbol{S}}_{x1,11} - \widehat{\boldsymbol{S}}_{x0,01} - (\widehat{\boldsymbol{S}}_{xx,11} - \widehat{\boldsymbol{S}}_{xx,01})\boldsymbol{\gamma}_{1c}\} \\ &\quad -(\pi_c \boldsymbol{S}_{xx,c})^{-1}\{\widehat{\boldsymbol{S}}_{x0,00} - \widehat{\boldsymbol{S}}_{x1,10} - (\widehat{\boldsymbol{S}}_{xx,00} - \widehat{\boldsymbol{S}}_{xx,10})\boldsymbol{\gamma}_{0c}\},\end{aligned}$$

third we observed that the difference between $\widehat{\boldsymbol{\beta}}_{c,\text{RI}} - \boldsymbol{\beta}_c$ and Δ_c has higher order following the same argument as (A.4). Therefore, we need only to find the asymptotic distribution of Δ_c.

Simple algebra gives

$$
\begin{aligned}
\Delta_c &= (\pi_c \boldsymbol{S}_{xx,c})^{-1}\Big[\frac{1}{n_1}\sum_{i=1}^{n} T_i D_i \boldsymbol{X}_i Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i) D_i \boldsymbol{X}_i Y_i(0) \\
&\quad -\frac{1}{n_1}\sum_{i=1}^{n} T_i D_i \boldsymbol{X}_i \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c1} + \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i) D_i \boldsymbol{X}_i \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c1} \\
&\quad -\frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)(1-D_i)\boldsymbol{X}_i Y_i(0) + \frac{1}{n_1}\sum_{i=1}^{n} T_i(1-D_i)\boldsymbol{X}_i Y_i(1) \\
&\quad +\frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)(1-D_i)\boldsymbol{X}_i \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c0} - \frac{1}{n_1}\sum_{i=1}^{n} T_i(1-D_i)\boldsymbol{X}_i \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c0}\Big] \\
&= (\pi_c \boldsymbol{S}_{xx,c})^{-1}\Big[\frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=a)} \boldsymbol{X}_i Y_i(1) + \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=c)} \boldsymbol{X}_i Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i) I_{(U_i=a)} \boldsymbol{X}_i Y_i(0) \\
&\quad -\frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=a)} \boldsymbol{X}_i \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c1} - \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=c)} \boldsymbol{X}_i \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c1} + \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i) I_{(U_i=a)} \boldsymbol{X}_i \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c1} \\
&\quad -\frac{1}{n_0}\sum_{i=1}^{n}(1-T_i) I_{(U_i=n)} \boldsymbol{X}_i Y_i(0) - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i) I_{(U_i=c)} \boldsymbol{X}_i Y_i(0) + \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=n)} \boldsymbol{X}_i Y_i(1) \\
&\quad +\frac{1}{n_0}\sum_{i=1}^{n}(1-T_i) I_{(U_i=n)} \boldsymbol{X}_i \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c0} + \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i) I_{(U_i=c)} \boldsymbol{X}_i \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c0} - \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=n)} \boldsymbol{X}_i \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c0}\Big] \\
&= (\pi_c \boldsymbol{S}_{xx,c})^{-1}\Big\{\frac{1}{n_1}\sum_{i=1}^{n} T_i X_i \left[I_{(U_i=a)}(Y_i(1) - \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c1}) + I_{(U_i=n)}(Y_i(1) - \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c0}) + I_{(U_i=c)}(Y_i(1) - \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c1})\right] \\
&\quad -\frac{1}{n_0}\sum_{i=1}^{n}(1-T_i) X_i \left[I_{(U_i=a)}(Y_i(0) - \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c1}) + I_{(U_i=n)}(Y_i(0) - \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c0}) + I_{(U_i=c)}(Y_i(0) - \boldsymbol{X}_i^{\tau}\boldsymbol{\gamma}_{c0})\right]\Big\}.
\end{aligned}
$$

According to the definitions of the residual potential outcomes $e_i'(1)$ and $e_i'(0)$ in the main text, the above formula reduces to

$$
\widetilde{\beta}_{c,\mathrm{RI}} - \beta_c = (\pi_c \boldsymbol{S}_{xx,c})^{-1}\left[\frac{1}{n_1}\sum_{i=1}^{n} T_i \boldsymbol{X}_i e_i'(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)\boldsymbol{X}_i e_i'(0)\right]. \tag{A.14}
$$

The representation in (A.14) implies the asymptotic covariance matrix according to Theorem 1 and the asymptotic normality of $\widetilde{\boldsymbol{\beta}}_{c,\mathrm{RI}}$ according to the FPCLT. □

Theorem 7: Behavior of $\widehat{\boldsymbol{\beta}}_{\mathbf{TSLS}}$. While the amount of notation and matrix algebra is considerably more in scope, the overall structure of the proof follows the earlier one for the OLS estimator for the ITT. In particular, we show the estimator asymptotically converges to a more tractable version that has a fixed portion, and then use the usual covariance argument on the remaining

terms. Before doing this, we first show the probability limits of the estimator by working through the matrix algebra.

Proof of Theorem 7. First, we find the probability limits of the TSLS estimators:

$$\begin{aligned}
\begin{pmatrix}\widehat{\boldsymbol{\gamma}}_{\text{TSLS}}\\ \widehat{\boldsymbol{\beta}}_{\text{TSLS}}\end{pmatrix} &= \left\{\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\boldsymbol{X}_i\\ T_i\boldsymbol{X}_i\end{pmatrix}(\boldsymbol{X}_i^{\intercal}, D_i\boldsymbol{X}_i^{\intercal})\right\}^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\boldsymbol{X}_i\\ T_i\boldsymbol{X}_i\end{pmatrix}Y_i^{\text{obs}}\right\}\\
&= \begin{pmatrix} n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal} & n^{-1}\sum_{i=1}^{n}D_i\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal}\\ n^{-1}\sum_{i=1}^{n}T_i\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal} & n^{-1}\sum_{i=1}^{n}T_iD_i\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal}\end{pmatrix}^{-1}\begin{pmatrix} n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_iY_i^{\text{obs}}\\ n^{-1}\sum_{i=1}^{n}T_i\boldsymbol{X}_iY_i^{\text{obs}}\end{pmatrix}\\
&\xrightarrow{P} \begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{G}\\ \boldsymbol{H}\end{pmatrix}. \qquad \text{(A.15)}
\end{aligned}$$

The above term $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{S}_{xx}$, and terms $(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{G}, \boldsymbol{H})$ are the population limits of the sample quantities. We will find each of them. Term $\boldsymbol{B}$ is

$$\begin{aligned}
\boldsymbol{B} &= E\left\{\frac{1}{n}\sum_{i=1}^{n}D_i\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal}\right\} = E\left\{\frac{1}{n}\sum_{i=1}^{n}T_iD_i\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal} + \frac{1}{n}\sum_{i=1}^{n}(1-T_i)D_i\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal}\right\}\\
&= E\left\{\frac{1}{n}\sum_{i=1}^{n}T_iI_{(U_i=a)}\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal} + \frac{1}{n}\sum_{i=1}^{n}T_iI_{(U_i=c)}\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal} + \frac{1}{n}\sum_{i=1}^{n}(1-T_i)I_{(U_i=a)}\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal}\right\}\\
&= p_1\pi_a\boldsymbol{S}_{xx,a} + p_1\pi_c\boldsymbol{S}_{xx,c} + p_0\pi_a\boldsymbol{S}_{xx,a}\\
&= \pi_a\boldsymbol{S}_{xx,a} + p_1\pi_c\boldsymbol{S}_{xx,c}.
\end{aligned}$$

Term $\boldsymbol{C}$ is $\boldsymbol{C} = E\left\{n^{-1}\sum_{i=1}^{n}T_i\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal}\right\} = p_1\boldsymbol{S}_{xx}$. Term $\boldsymbol{D}$ is

$$\begin{aligned}
\boldsymbol{D} &= E\left\{\frac{1}{n}\sum_{i=1}^{n}T_iD_i\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal}\right\} = E\left\{\frac{1}{n}\sum_{i=1}^{n}T_iI_{(U_i=a)}\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal} + \frac{1}{n}\sum_{i=1}^{n}T_iI_{(U_i=c)}\boldsymbol{X}_i\boldsymbol{X}_i^{\intercal}\right\}\\
&= p_1\pi_a\boldsymbol{S}_{xx,a} + p_1\pi_c\boldsymbol{S}_{xx,c}.
\end{aligned}$$

Term $\boldsymbol{G}$ is

$$\boldsymbol{G} = E\left\{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_iY_i^{\text{obs}}\right\} = E\left\{\frac{1}{n}\sum_{i=1}^{n}T_i\boldsymbol{X}_iY_i^{\text{obs}} + \frac{1}{n}\sum_{i=1}^{n}(1-T_i)\boldsymbol{X}_iY_i^{\text{obs}}\right\} = p_1\boldsymbol{S}_{x1} + p_0\boldsymbol{S}_{x0}.$$

Term $\boldsymbol{H}$ is $\boldsymbol{H} = E\left\{n^{-1}\sum_{i=1}^{n}T_i\boldsymbol{X}_iY_i^{\text{obs}}\right\} = p_1\boldsymbol{S}_{x1}$. We apply the following formula for the inverse of a block matrix:

$$\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{pmatrix}^{-1} = \begin{pmatrix}\boldsymbol{S}_{\boldsymbol{D}}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{S}_{\boldsymbol{A}}^{-1}\\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{S}_{\boldsymbol{D}}^{-1} & \boldsymbol{S}_{\boldsymbol{A}}^{-1}\end{pmatrix},$$

where $\boldsymbol{S}_{\boldsymbol{D}} = \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$ and $\boldsymbol{S}_{\boldsymbol{A}} = \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}$ are the Schur complements of blocks $\boldsymbol{D}$ and $\boldsymbol{A}$. Omitting some tedious matrix algebra, we obtain

$$\boldsymbol{S}_{\boldsymbol{D}} = p_0\pi_c\boldsymbol{S}_{xx,c}(\pi_a\boldsymbol{S}_{xx,a} + \pi_c\boldsymbol{S}_{xx,c})^{-1}\boldsymbol{S}_{xx}, \quad \boldsymbol{S}_{\boldsymbol{A}} = p_1p_0\pi_c\boldsymbol{S}_{xx,c},$$

and the inverse of the block matrix is

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx}^{-1}(\pi_a\boldsymbol{S}_{xx,a}+\pi_c\boldsymbol{S}_{xx,c})\boldsymbol{S}_{xx,c}^{-1} & -p_1^{-1}p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx}^{-1}(\pi_a\boldsymbol{S}_{xx,a}+p_1\pi_c\boldsymbol{S}_{xx,c})\boldsymbol{S}_{xx,c}^{-1} \\ -p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx,c}^{-1} & p_1^{-1}p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx,c}^{-1} \end{pmatrix}.$$

Therefore, according to (A.15), the probability limit of $\widehat{\boldsymbol{\gamma}}_{\text{TSLS}}$ is

$$\begin{aligned} & p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx}^{-1}(\pi_a\boldsymbol{S}_{xx,a}+\pi_c\boldsymbol{S}_{xx,c})\boldsymbol{S}_{xx,c}^{-1}(p_1\boldsymbol{S}_{x1}+p_0\boldsymbol{S}_{x0}) - p_1^{-1}p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx}^{-1}(\pi_a\boldsymbol{S}_{xx,a}+p_1\pi_c\boldsymbol{S}_{xx,c})\boldsymbol{S}_{xx,c}^{-1}(p_1\boldsymbol{S}_{x1}) \\ = \;& \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x0} - \pi_a\pi_c^{-1}\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xx,a}\boldsymbol{S}_{xx,c}^{-1}(\boldsymbol{S}_{x1}-\boldsymbol{S}_{x0}) \\ = \;& \boldsymbol{\gamma}_0 - \pi_a\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xx,a}\boldsymbol{\beta}_c \equiv \boldsymbol{\gamma}_\infty, \end{aligned} \tag{A.16}$$

and the probability limit of $\widehat{\boldsymbol{\beta}}_{\text{TSLS}}$ is

$$-p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx,c}^{-1}(p_1\boldsymbol{S}_{x1}+p_0\boldsymbol{S}_{x0}) + p_1^{-1}p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx,c}^{-1}(p_1\boldsymbol{S}_{x1}) = \pi_c^{-1}\boldsymbol{S}_{xx,c}^{-1}(\boldsymbol{S}_{x1}-\boldsymbol{S}_{x0}) = \boldsymbol{\beta}_c, \tag{A.17}$$

where we use $\boldsymbol{S}_{x1}-\boldsymbol{S}_{x0} = \pi_c(\boldsymbol{S}_{x1,c}-\boldsymbol{S}_{x0,c})$, which is guaranteed by exclusion restrictions.

We next find the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\text{TSLS}}$. Following the derivation in Corollary 3, we first write

$$\begin{pmatrix} \widehat{\boldsymbol{\gamma}}_{\text{TSLS}} \\ \widehat{\boldsymbol{\beta}}_{\text{TSLS}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\gamma}_\infty \\ \boldsymbol{\beta}_c \end{pmatrix} = \left\{ \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \boldsymbol{X}_i \\ T_i\boldsymbol{X}_i \end{pmatrix} (\boldsymbol{X}_i^\intercal, D_i\boldsymbol{X}_i^\intercal) \right\}^{-1} \left\{ \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \boldsymbol{X}_i(Y_i^{\text{obs}} - \boldsymbol{X}_i^\intercal\boldsymbol{\gamma}_\infty - D_i\boldsymbol{X}_i^\intercal\boldsymbol{\beta}_c) \\ T_i\boldsymbol{X}_i(Y_i^{\text{obs}} - \boldsymbol{X}_i^\intercal\boldsymbol{\gamma}_\infty - D_i\boldsymbol{X}_i^\intercal\boldsymbol{\beta}_c) \end{pmatrix} \right\},$$

then introduce

$$\begin{aligned} \Delta_{\text{TSLS}} &= \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} \left\{ \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \boldsymbol{X}_i(Y_i^{\text{obs}} - \boldsymbol{X}_i^\intercal\boldsymbol{\gamma}_\infty - D_i\boldsymbol{X}_i^\intercal\boldsymbol{\beta}_c) \\ T_i\boldsymbol{X}_i(Y_i^{\text{obs}} - \boldsymbol{X}_i^\intercal\boldsymbol{\gamma}_\infty - D_i\boldsymbol{X}_i^\intercal\boldsymbol{\beta}_c) \end{pmatrix} \right\} \\ &= \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} \begin{pmatrix} n^{-1}\sum_{i=1}^n T_i\boldsymbol{X}_i e_i''(1) + n^{-1}\sum_{i=1}^n (1-T_i)\boldsymbol{X}_i e_i''(0) \\ n^{-1}\sum_{i=1}^n T_i\boldsymbol{X}_i e_i''(1) \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} \begin{pmatrix} n^{-1}\sum_{i=1}^n T_i\boldsymbol{X}_i\{e_i''(1)-e_i''(0)\} + n^{-1}\sum_{i=1}^n \boldsymbol{X}_i e_i''(0) \\ n^{-1}\sum_{i=1}^n T_i\boldsymbol{X}_i e_i''(1) \end{pmatrix}, \end{aligned} \tag{A.18}$$

with $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ defined in (A.15) and $\{e_i''(1), e_i''(0)\}$ defined in Theorem 7, and finally recognize that the difference between the above two formulas has high order. Again we need only to find the asymptotic distribution of Δ_{TSLS}. The covariance of the second term on the right hand side of (A.18) is (dropping the constant sum of $\boldsymbol{X}_i e''(0)$)

$$\begin{aligned} & \text{cov}\begin{pmatrix} n^{-1}\sum_{i=1}^n T_i\boldsymbol{X}_i\{e_i''(1)-e_i''(0)\} \\ n^{-1}\sum_{i=1}^n T_i\boldsymbol{X}_i e_i''(1) \end{pmatrix} \\ = \;& \frac{1}{n^2}\frac{n_1 n_0}{n}\begin{pmatrix} \mathcal{S}(\boldsymbol{X}\varepsilon) & \frac{1}{2}[\mathcal{S}\{\boldsymbol{X}e''(1)\} - \mathcal{S}\{\boldsymbol{X}e''(0)\} + \mathcal{S}(\boldsymbol{X}\varepsilon)] \\ \frac{1}{2}[\mathcal{S}\{\boldsymbol{X}e''(1)\} - \mathcal{S}\{\boldsymbol{X}e''(0)\} + \mathcal{S}(\boldsymbol{X}\varepsilon)] & \mathcal{S}\{\boldsymbol{X}e''(1)\} \end{pmatrix}, \end{aligned}$$

where the off-diagonal term comes from the finite population covariance between $\boldsymbol{X}\{e''(1)-e''(0)\}$ and $\boldsymbol{X}e''(1)$. Therefore, according to (A.18), the asymptotic covariance of Δ_{TSLS} is the $(2,2)$ block of the following matrix

$$\frac{1}{n^2}\frac{n_1 n_0}{n}\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{pmatrix}^{-1}\cdot \begin{pmatrix}\mathcal{S}(\boldsymbol{X}\varepsilon) & \frac{1}{2}[\mathcal{S}\{\boldsymbol{X}e''(1)\}-\mathcal{S}\{\boldsymbol{X}e''(0)\}+\mathcal{S}(\boldsymbol{X}\varepsilon)]\\ \frac{1}{2}[\mathcal{S}\{\boldsymbol{X}e''(1)\}-\mathcal{S}\{\boldsymbol{X}e''(0)\}+\mathcal{S}(\boldsymbol{X}\varepsilon)] & \mathcal{S}\{\boldsymbol{X}e''(1)\}\end{pmatrix}\cdot\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{pmatrix}^{-\intercal},$$

which is

$$\begin{aligned}&\frac{1}{n^2}\frac{n_1 n_0}{n}\Big\{(p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx,c}^{-1})\mathcal{S}(\boldsymbol{X}\varepsilon)(p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx,c}^{-1})^{\intercal}+(p_1^{-1}p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx,c}^{-1})\mathcal{S}\{\boldsymbol{X}e''(1)\}(p_1^{-1}p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx,c}^{-1})^{\intercal}\\ &-(p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx,c}^{-1})[\mathcal{S}\{\boldsymbol{X}e''(1)\}-\mathcal{S}\{\boldsymbol{X}e''(0)\}+\mathcal{S}(\boldsymbol{X}\varepsilon)](p_1^{-1}p_0^{-1}\pi_c^{-1}\boldsymbol{S}_{xx,c}^{-1})^{\intercal}\Big\}\\ =\;&(\pi_c\boldsymbol{S}_{xx,c})^{-1}\left[\frac{\mathcal{S}\{\boldsymbol{X}e''(1)\}}{n_1}+\frac{\mathcal{S}\{\boldsymbol{X}e''(0)\}}{n_0}-\frac{\mathcal{S}(\boldsymbol{X}\varepsilon)}{n}\right](\pi_c\boldsymbol{S}_{xx,c})^{-1}.\end{aligned}$$

The asymptotic normality follows from the representation in (A.18) and the FPCLT. □

Theorem 8: Decomposition of variation in non-compliance. The following proof uses two facts: $\tau_a=\tau_n=0$, and $\tau=\pi_c\tau_c$.

Proof of Theorem 8. Write the total treatment effect variation as

$$\begin{aligned}S_{\tau\tau} &= \frac{1}{n}\sum_{i=1}^{n}(\tau_i-\tau)^2=\frac{1}{n}\sum_{i=1}^{n}\tau_i^2-\tau^2\\ &= \frac{1}{n}\sum_{i=1}^{n}I_{(U_i=c)}\tau_i^2-\pi_c^2\tau_c^2=\pi_c\left(\frac{1}{n_c}\sum_{i=1}^{n}I_{(U_i=c)}\tau_i^2-\tau_c^2\right)+\pi_c(1-\pi_c)\tau_c^2,\end{aligned}$$

the treatment effect variation explained by compliance status as

$$\begin{aligned}S_{\tau\tau,U} &= \sum_{u=c,a,n}\pi_u(\tau_u-\tau)^2=\pi_c(\tau_c-\pi_c\tau_c)^2+\pi_a(0-\pi_c\tau_c)^2+\pi_n(0-\pi_c\tau_c)^2\\ &= \pi_c\tau_c^2\left\{(1-\pi_c)^2+\pi_c(\pi_a+\pi_n)\right\}=\pi_c(1-\pi_c)\tau_c^2,\end{aligned}$$

and the subtotal treatment effect variation for compliers as

$$S_{\tau\tau,c}=\frac{1}{n_c}\sum_{i=1}^{n}I_{(U_i=c)}(\tau_i-\tau_c)^2=\frac{1}{n_c}\sum_{i=1}^{n}I_{(U_i=c)}\tau_i^2-\tau_c^2.$$

Therefore, the above three terms has the relationship $S_{\tau\tau}=\pi_c S_{\tau\tau,c}+S_{\tau\tau,U}$.

The decomposition $S_{\tau\tau,c}=S_{\delta\delta,c}+S_{\varepsilon\varepsilon,c}$ follows immediately from the definition of $\boldsymbol{\beta}_c$. □

Appendix B More detailed comments

Appendices B.1–B.5 give more details of some technical issues and extensions mentioned in the main text, and Appendix B.6 contains the proofs of the results in Appendix B.

Appendix B.1 Covariate adjustment to improve efficiency

In the main text, the role of covariates has been to model the treatment effect alone. In general, we also want to use covariates to reduce sampling variability of $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$, just as we can use covariates to get more precise estimates of the average treatment effect. In particular, the goal is to more precisely estimate $\widehat{\boldsymbol{S}}_{xt} \in \mathbb{R}^K$; because these are the only random components in $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$, if we estimate them more precisely, we estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ more precisely as well. Let $\boldsymbol{W}_i \in \mathbb{R}^J$ denote a vector of pretreatment covariates without the intercept term. Because $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ have different roles in estimation, they may also contain different sets of covariates, though, in practice, $\boldsymbol{X}$ is likely to be a subset of $\boldsymbol{W}$.

Following the covariate adjustment approach in survey sampling, we can obtain a model-assisted estimator for $\boldsymbol{\beta}$ that uses $\boldsymbol{W}$ to reduce sampling variability. To see this, we need several definitions. Define $\overline{\boldsymbol{W}} = n^{-1}\sum_{i=1}^n \boldsymbol{W}_i$ and $\boldsymbol{S}_{ww} = n^{-1}\sum_{i=1}^n \boldsymbol{W}_i\boldsymbol{W}_i^{\intercal}$, with $\det(\boldsymbol{S}_{ww}) > 0$; define $\overline{\boldsymbol{W}}_t$ and $\widehat{\boldsymbol{S}}_{ww,t}$ as the sample mean and covariance of $\boldsymbol{W}$ under treatment arm t; define $\widehat{\boldsymbol{B}}_t \in \mathbb{R}^{J\times K}$ as the regression coefficient of $Y^{\mathrm{obs}}\boldsymbol{X}$ on $\boldsymbol{W}$ for treatment arm t:

$$\widehat{\boldsymbol{B}}_t = \widehat{\boldsymbol{S}}_{ww,t}^{-1}\left\{\frac{1}{n_t}\sum_{i=1}^n I_{(T_i=t)}\boldsymbol{W}_i(Y_i^{\mathrm{obs}}\boldsymbol{X}_i)^{\intercal}\right\}.$$

The model-assisted estimator for $\boldsymbol{S}_{xt}$ is then

$$\widehat{\boldsymbol{S}}_{xt}^{w} = \widehat{\boldsymbol{S}}_{xt} - \widehat{\boldsymbol{B}}_t^{\intercal}(\bar{\boldsymbol{W}}_t - \bar{\boldsymbol{W}}), \quad (t = 0, 1).$$

As a result, we can improve the randomization-based estimator by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}^{w} = \boldsymbol{S}_{xx}^{-1}(\widehat{\boldsymbol{S}}_{x1}^{w} - \widehat{\boldsymbol{S}}_{x0}^{w}).$$

Theorem A.1. The model-assisted estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}^{w}$ is consistent for $\boldsymbol{\beta}$ with asymptotic covariance

$$\boldsymbol{S}_{xx}^{-1}\left[\frac{\mathcal{S}\{\boldsymbol{E}(1)\}}{n_1} + \frac{\mathcal{S}\{\boldsymbol{E}(0)\}}{n_0} - \frac{\mathcal{S}(\boldsymbol{\Delta})}{n}\right]\boldsymbol{S}_{xx}^{-1},$$

where $\boldsymbol{E}_i(t) = Y_i(t)\boldsymbol{X}_i - \boldsymbol{B}_t^{\intercal}(\boldsymbol{W}_i - \bar{\boldsymbol{W}})$ is the residual term and $\boldsymbol{\Delta}_i = \boldsymbol{E}_i(1) - \boldsymbol{E}_i(0)$.

The estimator, $\widehat{\boldsymbol{\beta}}^{w}_{\mathrm{RI}}$ uses covariates both to estimate treatment effect variation and to reduce sampling variability. Asymptotically, as long as $\boldsymbol{W}$ is predictive of the marginal potential outcomes, the model-assisted estimator will improve precision over the unassisted estimators.

Appendix B.2 Fisherian exact inference

When $\varepsilon_i = 0$ for all i, we can obtain exact inference for $\boldsymbol{\beta}$ based on the Fisher randomization test (Rubin, 1980; Rosenbaum, 2002; Ding et al., 2016). With a known $\boldsymbol{\beta}$, the null hypothesis

$$H_0(\boldsymbol{\beta}) : Y_i(1) - Y_i(0) = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} \text{ for all } i \tag{A.19}$$

is sharp in the sense of allowing for full imputation of all missing potential outcomes based on the observed data. We can perform randomization test using any sensible test statistic measuring the deviation from the null hypothesis $H_0(\boldsymbol{\beta})$, for example, the test statistic $t(\boldsymbol{T}, \boldsymbol{Y}^{\mathrm{obs}}; \boldsymbol{\beta})$ can be the difference-in-means, difference-in-medians or the Kolmogorov–Smirnov statistics comparing two samples $\{Y_i^{\mathrm{obs}} - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} : T_i = 1, i = 1, \ldots, n\}$ and $\{Y_i^{\mathrm{obs}} : T_i = 0, i = 1, \ldots, n\}$. Then we can obtain a $(1-\alpha)$ level confidence region for $\boldsymbol{\beta}$ by inverting a sequence of randomization tests:

$$\mathrm{CR}_\alpha = \{\boldsymbol{\beta} : \text{Randomization test fails to reject } H_0(\boldsymbol{\beta}) \text{ at significance level } \alpha\}.$$

The confidence region CR_α is exact regardless of the sample size, and it is valid for general designs of experiments if we use the corresponding assignment mechanism to simulate the null distribution of the test statistic. Due to the duality between testing and interval estimation, we reject $H_0(\boldsymbol{X})$ with $\boldsymbol{\beta}_1 = 0$ in Section 3.3 if $\mathrm{CR}_\alpha \cap \{\boldsymbol{\beta} : \boldsymbol{\beta}_1 = 0\}$ is an empty set, which controls the type one error rate by α.

Appendix B.3 A Variance Ratio Test

Raudenbush and Bloom (2015) have noticed that if the variance of the treatment potential outcome is smaller than the control potential outcome, then the correlation between the individual treatment effect and the control potential outcome is negative. This statement does not involve any covariates, but it can be generalized to incorporate systematic and idiosyncratic treatment effect variation. Below we give a finite population version of their result.

Theorem A.2. If the finite population variance of $\{Y_i(1) - \boldsymbol{X}_i'\boldsymbol{\beta}\}_{i=1}^n$ is smaller than $\{Y_i(0)\}_{i=1}^n$, then the idiosyncratic treatment effect variation, $\{\varepsilon_i\}_{i=1}^n$, is negatively correlated with the control potential outcomes.

Because the condition in Theorem A.2 depends only on the marginal distributions of the potential outcomes, we propose a formal variance ratio test of it using the observed data, which is a generalization of a similar theorem in Ding et al. (2016):

Theorem A.3. The variance ratio test with rejection region

$$\frac{\log s_1^2 - \log s_0^2}{\sqrt{(\widehat{\kappa}_1 - 1)/n_1 + (\widehat{\kappa}_0 - 1)/n_0}} < \Phi^{-1}(\alpha),$$

has size at least as large as α, where s_1^2 and $\widehat{\kappa}_1$ are the sample variance and kurtosis of $\{Y_i^{\text{obs}} - \boldsymbol{X}_i^{\intercal}\widehat{\boldsymbol{\beta}}_{\text{RI}} : T_i = 1, i = 1, \ldots, n\}$, and s_0^2 and $\widehat{\kappa}_0$ are the sample variance and kurtosis of $\{Y_i^{\text{obs}} : T_i = 0, i = 1, \ldots, n\}$, and $\Phi^{-1}(\alpha)$ is the α-th quantile of the standard normal distribution.

For finite population inference, the above test in Theorem A.3 is generally conservative, but for superpopulation inference, it is asymptotically exact.

Note that Raudenbush and Bloom (2015) and Theorem A.2 are only about detecting a negative association. Unfortunately, there is no testable condition for a positive association.

Appendix B.4 More on noncompliance: estimating the bounds of the R^2s

The component $S_{\tau\tau,U}$ and and the probability π_c are directly identifiable according to previous discussion. Furthermore, $S_{\delta\delta,c}$ is also identifiable according to the following result.

Corollary A.1. $S_{\delta\delta,c}$ can be expressed as the expectation of the following quantity:

$$\frac{1}{\pi_c}\left\{\frac{1}{n}\sum_{i=1}^{n}(\delta_i - \tau_c)^2 - \frac{1}{n_1}\sum_{i=1}^{n}T_i(1 - D_i)(\delta_i - \tau_c)^2 - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_i)D_i(\delta_i - \tau_c)^2\right\}.$$

Because π_c, $\delta_i = \boldsymbol{X}_i^{\intercal}\boldsymbol{\beta}_c$ and τ_c can be estimated by a plug-in approach, $S_{\delta\delta,c}$ can also be estimated from the observed data.

In the ITT case, estimation of the residual distributions are straightforward. In the noncompliance case, however, we need more discussion about the estimation of $F_{1c}(y)$ and $F_{0c}(y)$, because U_i is a latent variable. To avoid notational clatter, we assume that $\boldsymbol{\gamma}_{c1}$ and $\boldsymbol{\gamma}_{c0}$ are known; in practice we can replace them by the randomization-based estimators $\widehat{\boldsymbol{\gamma}}_{c1,\text{RI}}$ and $\widehat{\boldsymbol{\gamma}}_{c0,\text{RI}}$, and the consistency of the final estimator will not be affected. Recall the potential residuals $e_i'(1)$ and $e_i'(0)$ defined in (17), and its observed value $e_i' = T_i e_i'(1) + (1 - T_i)e_i'(0)$. We define the following quantities

$$\begin{aligned}
\widehat{F}_{11}(y) &= \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n} T_i D_i I_{(e_i' \le y)}, & \widehat{F}_{10}(y) &= \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n} T_i(1 - D_i) I_{(e_i' \le y)},\\
\widehat{F}_{01}(y) &= \tfrac{1}{n_0}\textstyle\sum_{i=1}^{n}(1 - T_i) D_i I_{(e_i' \le y)}, & \widehat{F}_{00}(y) &= \tfrac{1}{n_0}\textstyle\sum_{i=1}^{n}(1 - T_i)(1 - D_i) I_{(e_i' \le y)}.
\end{aligned} \tag{A.20}$$

Similar to Corollary 3, we have the following results.

Corollary A.2. For any y,

$$E\{\widehat{F}_{11}(y) - \widehat{F}_{01}(y)\} = \pi_c F_{1c}(y), \quad E\{\widehat{F}_{00}(y) - \widehat{F}_{10}(y)\} = \pi_c F_{0c}(y).$$

Therefore, we can estimate $F_{1c}(y)$ by $\{\widehat{F}_{11}(y) - \widehat{F}_{01}(y)\}/\widehat{\pi}_c$, and estimate $F_{0c}(y)$ by $\{\widehat{F}_{00}(y) - \widehat{F}_{10}(y)\}/\widehat{\pi}_c$. As we mentioned before, in practice, we use $\widehat{e}_i'$ instead of e_i' in the formulas in (A.20).

Appendix B.5 Proofs of the theorems and corollaries in Appendix B

Proof of Theorem A.1. The population-level OLS regression matrix of $Y(t)\boldsymbol{X}$ onto $\boldsymbol{W}$ is

$$\boldsymbol{B}_t = \boldsymbol{S}_{ww}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{W}_i \{Y_i(t)\boldsymbol{X}_i\}^{\intercal} \right\} \in \mathbb{R}^{J \times K}.$$

Define $\widetilde{\boldsymbol{S}}_{xt}^{w} = \widehat{\boldsymbol{S}}_{xt} + \boldsymbol{B}_t^{\intercal}(\bar{\boldsymbol{W}} - \bar{\boldsymbol{W}}_t)$ and $\widetilde{\boldsymbol{\beta}}_{\mathrm{RI}}^{w} = \boldsymbol{S}_{xx}^{-1}(\widetilde{\boldsymbol{S}}_{x1}^{w} - \widetilde{\boldsymbol{S}}_{x0}^{w})$. According to the same argument as (A.4), $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ and $\widetilde{\boldsymbol{\beta}}_{\mathrm{RI}}^{w}$ have the same asymptotic covariance, and in the following we need only to discuss the covariance of $\widetilde{\boldsymbol{\beta}}_{\mathrm{RI}}^{w}$. Because

$$\begin{aligned} \widetilde{\boldsymbol{S}}_{x1}^{w} - \widetilde{\boldsymbol{S}}_{x0}^{w} &= \frac{1}{n_1} \sum_{i=1}^{n} T_i \left\{ Y_i(1)\boldsymbol{X}_i + \boldsymbol{B}_1^{\intercal}(\bar{\boldsymbol{W}} - \boldsymbol{W}_i) \right\} - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) \left\{ Y_i(0)\boldsymbol{X}_i + \boldsymbol{B}_0^{\intercal}(\bar{\boldsymbol{W}} - \boldsymbol{W}_i) \right\} \\ &= \frac{1}{n_1} \sum_{i=1}^{n} T_i \boldsymbol{E}_i(1) - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) \boldsymbol{E}_i(0) \end{aligned}$$

can be represented as the difference between the sample means of $\boldsymbol{E}_i(1)$ and $\boldsymbol{E}_i(0)$, applying Theorem 2 we can obtain its covariance:

$$\mathrm{cov}\left(\widetilde{\boldsymbol{S}}_{x1}^{w} - \widetilde{\boldsymbol{S}}_{x0}^{w} \right) = \frac{\mathcal{S}\{\boldsymbol{E}(1)\}}{n_1} + \frac{\mathcal{S}\{\boldsymbol{E}(0)\}}{n_0} - \frac{\mathcal{S}\{\boldsymbol{\Delta}\}}{n},$$

which completes the proof. □

Proof of Theorem A.2. For simplicity, we abuse the variance and covariance notation for finite population. For example, $\mathrm{var}\{Y(0)\} = \sum_{i=1}^{n} \{Y_i(0) - \bar{Y}(0)\}^2/(n-1)$. If $\mathrm{var}\{Y(1) - \boldsymbol{X}^{\intercal}\boldsymbol{\beta}\} \leq \mathrm{var}\{Y(0)\}$, then $\mathrm{var}\{Y(0) + \varepsilon\} \leq \mathrm{var}\{Y(0)\}$. Expanding the left hand side,

$$\mathrm{var}\{Y(0)\} + \mathrm{var}\{\varepsilon\} + 2\mathrm{cov}\{Y(0), \varepsilon\} \leq \mathrm{var}\{Y(0)\},$$

which implies $2\mathrm{cov}\{Y(0), \varepsilon\} \leq -\mathrm{var}\{\varepsilon\} < 0$. □

Although it is straightforward to prove the conclusion for super population inference of Theorem A.3 by using Ding et al. (2016, Theorem 2, Supplementary Material) and Slutsky's Theorem, it is less obvious to prove the conclusion for finite population inference. To simplify the proof, we first prove the following lemma. Let $(c_1, \cdots, c_n)^{\intercal}$ and $(d_1, \ldots, d_n)^{\intercal}$ be two vectors of nonnegative constants with the same mean $m > 0$ but different variances S_{cc} and S_{dd}. The difference vector $(c_1 - d_1, \ldots, c_n - d_n)^{\intercal}$ has mean zero and variance $S_{c-d,c-d}$. Let

$$\widehat{\theta}_c = \frac{1}{n_1}\sum_{i=1}^{n} T_i c_i, \quad \widehat{\theta}_d = \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)d_i$$

be two sample means of the treatment and control group, respectively.

Lemma A.5. Under the regularity conditions for the FPCLT, $\log\widehat{\theta}_c - \log\widehat{\theta}_d$ has asymptotic mean zero and variance

$$\frac{1}{m^2}\left(\frac{S_{cc}}{n_1} + \frac{S_{dd}}{n_0} - \frac{S_{c-d,c-d}}{n}\right). \tag{A.21}$$

Proof of Lemma A.5. According to the FPCLT, we have the following joint asymptotic normality of $\widehat{\theta}_c$ and $\widehat{\theta}_d$:

$$\begin{pmatrix}\widehat{\theta}_c \\ \widehat{\theta}_d\end{pmatrix} = \begin{pmatrix} n_1^{-1}\sum_{i=1}^{n} T_i c_i \\ n_0^{-1}\sum_{i=1}^{n}(1-T_i)d_i\end{pmatrix} \overset{a}{\sim} N\left[\begin{pmatrix}m\\m\end{pmatrix}, \begin{pmatrix}V_{cc} & V_{cd}\\ V_{cd} & V_{dd}\end{pmatrix}\right],$$

where

$$V_{cc} = \frac{n_0}{n_1 n}S_{cc}, \quad V_{dd} = \frac{n_1}{n_0 n}S_{dd}, \quad V_{cd} = -\frac{1}{2n}(S_{cc} + S_{dd} - S_{c-d,c-d}).$$

Applying Taylor expansion at m, we have $\log\widehat{\theta}_c - \log\widehat{\theta}_d = \{(\widehat{\theta}_c - m) - (\widehat{\theta}_d - m)\}/m + o_P(n^{-1/2})$, which, coupled with Neyman (1923)'s variance formula, gives the asymptotic variance of $\log\widehat{\theta}_c - \log\widehat{\theta}_d$ in (A.21). □

Proof of Theorem A.3. First, as a direct consequence of Lemma A.5, the finite sample variance is always larger than the super population variance, unless $S_{c-d,c-d} = 0$. Therefore, we need only to show that the test in Theorem A.3 is asymptotically exact for super population inference, and the asymptotic size of the test is no larger than α for finite population inference.

Second, replacing $\boldsymbol{\beta}$ by its consistent estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ does not affect the asymptotic distribution of the test statistic, due to Slutsky's Theorem. For simplicity, we treat $\boldsymbol{\beta}$ as known in our asymptotic analysis.

With the two ingredients above, Theorem A.3 follows directly from the variance ratio test in Ding et al. (2016, Theorem 2, Supplementary Material). □

Proof of Corollary A.1. The conclusion follows from

$$\begin{aligned} E\left\{\frac{1}{n_1}\sum_{i=1}^{n} T_i(1-D_i)(\delta_i-\tau_c)^2\right\} &= E\left\{\frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=n)}(\delta_i-\tau_c)^2\right\} = \frac{1}{n}\sum_{i=1}^{n} I_{(U_i=n)}(\delta_i-\tau_c)^2, \\ E\left\{\frac{1}{n_0}\sum_{i=1}^{n} (1-T_i)D_i(\delta_i-\tau_c)^2\right\} &= E\left\{\frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) I_{(U_i=a)}(\delta_i-\tau_c)^2\right\} = \frac{1}{n}\sum_{i=1}^{n} I_{(U_i=a)}(\delta_i-\tau_c)^2. \end{aligned}$$

□

Proof of Corollary A.2. We rewrite

$$\begin{aligned} \widehat{F}_{11}(y) &= \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=c)} I_{\{e_i(1)\le y\}} + \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=a)} I_{\{e_i(1)\le y\}}, \\ \widehat{F}_{10}(y) &= \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=n)} I_{\{e_i(1)\le y\}}, \\ \widehat{F}_{01}(y) &= \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) I_{(U_i=a)} I_{\{e_i(0)\le y\}}, \\ \widehat{F}_{00}(y) &= \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) I_{(U_i=c)} I_{\{e_i(0)\le y\}} + \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) I_{(U_i=n)} I_{\{e_i(0)\le y\}}. \end{aligned}$$

In the above formulas, the random components are the T_i's, and therefore, the corollary follows from Lemma A.1 and the linearity of expectations. □