Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://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 (2018): Decomposing Treatment Effect Variation, Journal of the American Statistical Association, DOI: 10.1080/01621459.2017.1407322

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

Decomposing Treatment Effect Variation *

Peng Ding
UC Berkeley

Avi Feller
UC Berkeley

Luke Miratrix
Harvard GSE

November 10, 2017

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 paper 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.

Key Words: Noncompliance; Heterogeneous treatment effect; Idiosyncratic treatment effect variation; Randomization inference; Systematic treatment effect variation.

*Peng Ding (Email: pengdingpku@berkeley.edu) is Assistant Professor, Department of Statistics, University of California, Berkeley. Avi Feller (Email: afeller@berkeley.edu) is Assistant Professor, Goldman School of Public Policy, University of California, Berkeley. Luke Miratrix (Email: lmiratrix@g.harvard.edu) is Assistant Professor, Harvard Graduate School of Education. We 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. We also thank our reviewers who helped us sharpen our mathematical presentations, in particular the asymptotic arguments. We 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).

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 papers on the topic in statistics and machine learning (Hill, 2011; Athey and Imbens, 2016; Wager and Athey, 2017), biostatistics (Huang et al., 2012; Matsouaka et al., 2014), education (Raudenbush and Bloom, 2015), economics (Heckman et al., 1997; Crump et al., 2008; Djebbari and Smith, 2008), political science (Green and Kern, 2012; Imai and Ratkovic, 2013), and other areas.

This paper 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 et al., 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 et al., 1996). We show that this estimator is nearly identical to the two-stage least squares 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 et al., 1997; Djebbari and Smith, 2008), again 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 et al. (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 $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{Ki})^\intercal \in \mathbb{R}^K$ 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 pre-treatment. 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 $\boldsymbol{T} = (T_1, \ldots, T_n)^\intercal$. 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 (1923)'s 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

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

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:

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

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 $\mathcal{S}\{\boldsymbol{V}(1)-\boldsymbol{V}(0)\}=\mathbf{0}_{K\times K}$. 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 analogues:

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

where

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

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 non-negative 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, e.g., 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

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

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

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

Following Heckman et al. (1997) and Djebbari and Smith (2008), we call $\delta_i=\boldsymbol{X}_i^{\intercal}\boldsymbol{\beta}$ the *systematic treatment effect variation* explained by the observed covariates, $\boldsymbol{X}_i$, and call ε_i the *idiosyncratic treatment effect variation* not explained by $\boldsymbol{X}_i$.

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

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

where $\boldsymbol{S}_{xx}$ is non-degenerate, 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 $\boldsymbol{S}_{x\varepsilon} = 0$, i.e., 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, $\delta_i = \boldsymbol{X}_i^{\intercal}\boldsymbol{\beta}$, 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), i.e. the ordinary least squares (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 $e_i(1) = Y_i(1) - \boldsymbol{X}_i^{\intercal}\boldsymbol{\gamma}_1$ and $e_i(0) = Y_i(0) - \boldsymbol{X}_i^{\intercal}\boldsymbol{\gamma}_0$ 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 analogues of $\boldsymbol{S}_{x1}$ and $\boldsymbol{S}_{x0}$:

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

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

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

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

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

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 formulae less cluttered by subscripts and excessive annotation. As with the vector version of Neyman's formula, the square root of the diagonal of $\mathrm{cov}(\widehat{\boldsymbol{\beta}}_{\mathrm{RI}})$ gives the standard errors of β.

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 $\mathrm{cov}(\widehat{\boldsymbol{\beta}}_{\mathrm{RI}})$ by

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

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 $\widehat{\tau} = \boldsymbol{X}_i^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ 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 paper, 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 non-zero limiting values; (iii) both $\boldsymbol{S}_{xx}$ and its limit have full rank K; (iv) there are no individual extreme values in the limit: $\max_{1\leq i\leq n} ||V_i - \bar{V}||_2^2/n \to 0$, 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 $1 \leq k, k' \leq K$ 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 (cf 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 et al. (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:

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

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

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

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 squares estimator under complete randomization, without assuming that model (8) is correctly specified. In particular, we do not assume any i.i.d. 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

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

where

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

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; $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ therefore follows an asymptotic normal distribution with mean $\boldsymbol{\beta}$ and covariance matrix

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

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

This estimate is simply the difference between $\widehat{\boldsymbol{\gamma}}_{1,\text{OLS}} = \widehat{\boldsymbol{S}}_{xx,1}^{-1}\widehat{\boldsymbol{S}}_{x1}$ and $\widehat{\boldsymbol{\gamma}}_{0,\text{OLS}} = \widehat{\boldsymbol{S}}_{xx,0}^{-1}\widehat{\boldsymbol{S}}_{x0}$, 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 $\widehat{e}_i = Y_i^{\text{obs}} - \boldsymbol{X}_i^{\intercal}\widehat{\boldsymbol{\gamma}}_{1,\text{OLS}}$, and for control units, define residual $\widehat{e}_i = Y_i^{\text{obs}} - \boldsymbol{X}_i^{\intercal}\widehat{\boldsymbol{\gamma}}_{0,\text{OLS}}$. 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 analogues, and conservatively estimate the asymptotic covariance matrix (9) by

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

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; Lin, 2013; Angrist and Pischke, 2008).

Importantly, $\widehat{\boldsymbol{\beta}}_{\text{RI}}$ and $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ are quite similar in form. In particular, $\widehat{\boldsymbol{\beta}}_{\text{RI}}$ uses the true $\boldsymbol{S}_{xx}$ while $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ 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.

The RI estimator, by comparison, has no adjustment whatsoever, and so cannot account for such random covariate imbalances. However, in Section 3.4 below 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. For related discussion, see Cochran (1977) on ratio estimators in surveys.

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 dependance 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 sub-vector of $\widehat{\boldsymbol{\beta}}$ and sub-matrix of $\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}})$, corresponding to the non-intercept coordinates of $\boldsymbol{X}$. We reject when

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

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 et al., 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 et al., 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:

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

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 $\{\widehat{\delta}_i = \boldsymbol{X}_i^\intercal \widehat{\boldsymbol{\beta}}\}_{i=1}^n$, where $\widehat{\boldsymbol{\beta}}$ is a consistent estimator, e.g., $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ or $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. 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

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

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 $\boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$, $F_{e1}(y)$ 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 (1923)'s estimate of overall ITT, $\widehat{\tau} = n_1^{-1}\sum_{i=1}^n T_i Y_i^{\mathrm{obs}} - n_0^{-1}\sum_{i=1}^n (1 - T_i) Y_i^{\mathrm{obs}}$, extending previous work by Heckman et al. (1997) and Aronow et al. (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,

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

where $S_{\tau\tau} = S_{\delta\delta} + S_{\varepsilon\varepsilon}$. As we discuss above, Neyman (1923) proposed a lower bound for the overall $\mathrm{var}(\widehat{\tau})$ under the assumption of a constant treatment effect, $S_{\tau\tau} = 0$. More recently, Aronow et al. (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 $\mathrm{var}(\widehat{\tau})$ 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 :

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

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 $(0,1)$ corresponds 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^2_τ.

Corollary 2. The sharp bounds on R^2_τ are

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

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

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

We estimate these bounds via plug-in estimates. Note that Djebbari and Smith (2008) explore 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 et al. (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, i.e., $\pi_d = 0$. 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 et al. (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 et al., 1995; Staiger and Stock, 1997), i.e., 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:

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

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

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

Then, analogous to (5),

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

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 et al., 1996; Abadie, 2003; Angrist and Pischke, 2008), we use estimates from observed subgroups to estimate the desired quantities of interest. Define sample moments:

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

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

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

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 squares 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:

$$e_i'(1) = \begin{cases} Y_i(1) - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\gamma}_{1c}, \\ Y_i(1) - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\gamma}_{0c}, \\ Y_i(1) - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\gamma}_{1c}, \end{cases} \quad e_i'(0) = \begin{cases} Y_i(0) - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\gamma}_{1c}, & U_i = a, \\ Y_i(0) - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\gamma}_{0c}, & U_i = n, \\ Y_i(0) - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\gamma}_{0c}, & U_i = c. \end{cases} \tag{17}$$

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 analogues; $\mathcal{S}(\varepsilon\boldsymbol{X})$, however, is unidentifiable. For units with $D_i = 1$, we define the residual $\widehat{e}_i' = Y_i^{\text{obs}} - \boldsymbol{X}_i^{\intercal}\widehat{\boldsymbol{\gamma}}_{c1,\text{RI}}$, and for units with $D_i = 0$, we define the residual $\widehat{e}_i' = Y_i^{\text{obs}} - \boldsymbol{X}_i^{\intercal}\widehat{\boldsymbol{\gamma}}_{c0,\text{RI}}$. Therefore, we can obtain a conservative estimate for the asymptotic covariance (16) by the following sandwich form:

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

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 squares (TSLS) setting in econometrics (e.g., Angrist and Pischke, 2008). First, we impose a linear regression model with treatment-covariate interactions:

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

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," i.e., 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 $(\widehat{\boldsymbol{\gamma}}_{\text{TSLS}}, \widehat{\boldsymbol{\beta}}_{\text{TSLS}})$ are the solutions to the following estimating equations:

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

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 $\widehat{\boldsymbol{\beta}}_{\text{TSLS}}$ is consistent for $\boldsymbol{\beta}_c$ across randomizations.

Theorem 7. Over all randomizations, the TSLS estimator $\widehat{\boldsymbol{\beta}}_{\text{TSLS}}$ 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 the residual potential outcomes are defined as

$$e_i''(1) = \begin{cases} Y_i(1) - \boldsymbol{X}_i^{\intercal}(\boldsymbol{\gamma}_\infty + \boldsymbol{\beta}_c), \\ Y_i(1) - \boldsymbol{X}_i^{\intercal}\boldsymbol{\gamma}_\infty, \\ Y_i(1) - \boldsymbol{X}_i^{\intercal}(\boldsymbol{\gamma}_\infty + \boldsymbol{\beta}_c), \end{cases} \quad e_i''(0) = \begin{cases} Y_i(0) - \boldsymbol{X}_i^{\intercal}(\boldsymbol{\gamma}_\infty + \boldsymbol{\beta}_c), & U_i = a, \\ Y_i(0) - \boldsymbol{X}_i^{\intercal}\boldsymbol{\gamma}_\infty, & U_i = n \\ Y_i(0) - \boldsymbol{X}_i^{\intercal}\boldsymbol{\gamma}_\infty, & U_i = c, \end{cases}$$

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

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

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

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

Finally, while $\widehat{\boldsymbol{\beta}}_{\text{TSLS}}$ is a consistent estimator for $\boldsymbol{\beta}_c$, $\widehat{\boldsymbol{\gamma}}_{\text{TSLS}}$ is not, in general, a consistent estimator for $\boldsymbol{\gamma}_{c0}$; that is, $\boldsymbol{\gamma}_\infty \neq \boldsymbol{\gamma}_{c0}$. Instead, $\widehat{\boldsymbol{\gamma}}_{\text{TSLS}}$ 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 $\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}})$ 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 $\widehat{\boldsymbol{\beta}}_{c,\text{RI}}$ or TSLS estimator $\widehat{\boldsymbol{\beta}}_{\text{TSLS}}$; 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, $\{e_i'(1) = Y_i(1) - \boldsymbol{X}_i^{\intercal}\boldsymbol{\gamma}_{1c} : U_i = c, i = 1,\ldots,n\}$ and $\{e_i'(0) = Y_i(0) - \boldsymbol{X}_i^{\intercal}\boldsymbol{\gamma}_{0c} : U_i = c, i = 1,\ldots,n\}$, 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

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

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 "pre-test"-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 shows the power of these tests, with $\alpha = 0.05$, 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, $n = 3,586$.

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 $\widehat{\boldsymbol{\beta}}_{c,\text{RI}}$ and $\widehat{\boldsymbol{\beta}}_{\text{TSLS}}$. 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

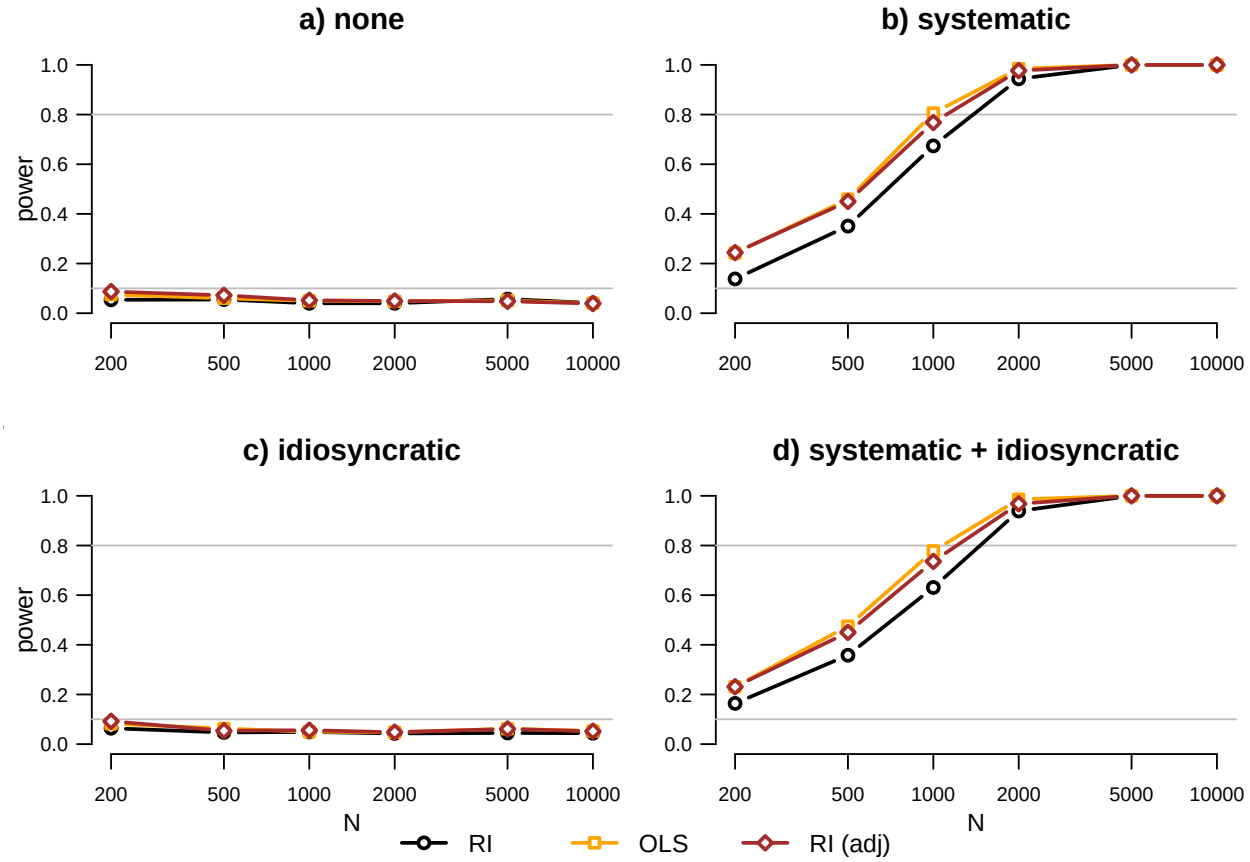


Figure 1: Power of the tests based on $\widehat{\boldsymbol{\beta}}_{\text{RI}}$, $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$, and $\widehat{\boldsymbol{\beta}}^{w}_{\text{RI}}$.

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 three- and four-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

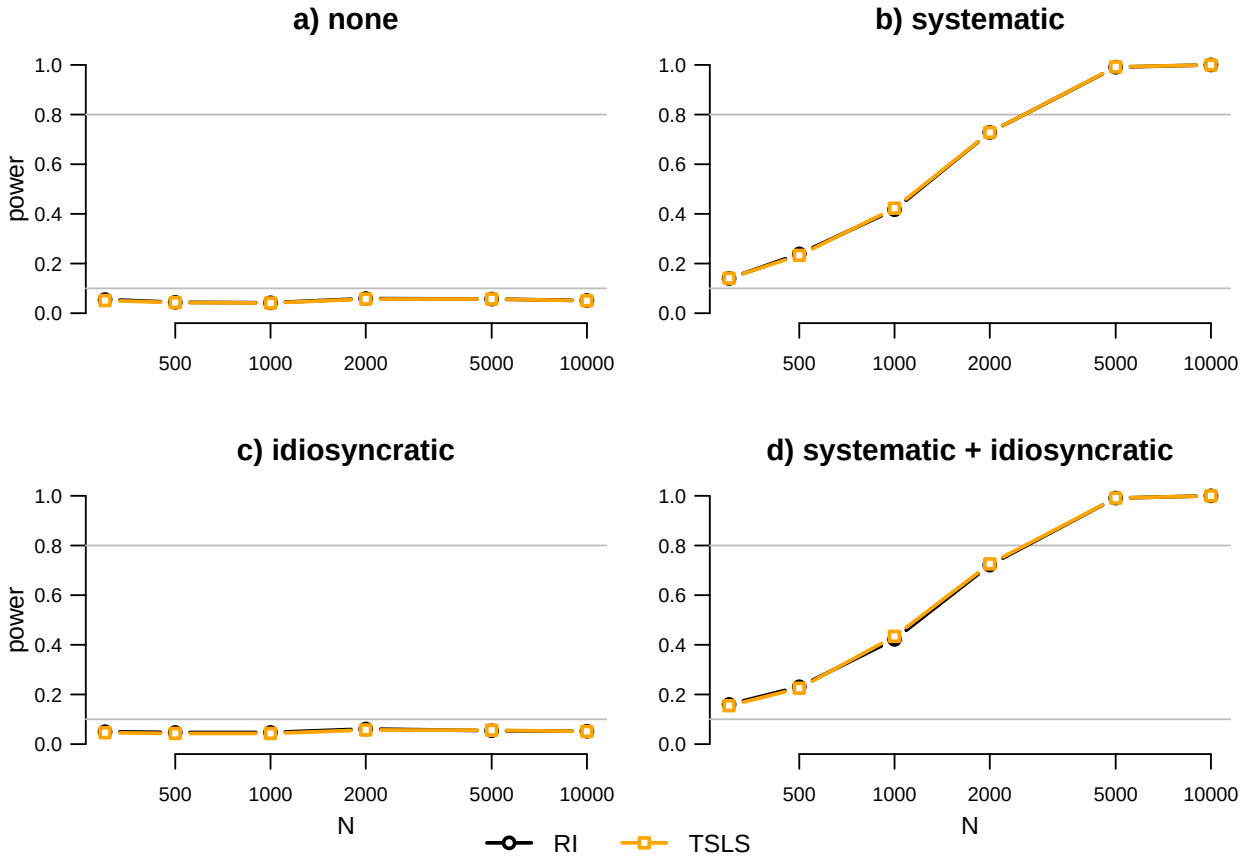


Figure 2: Power of the tests based on $\widehat{\boldsymbol{\beta}}_{c,\mathrm{RI}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{TSLS}}$.

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 papers argue that these topline results mask important treatment effect variation (e.g., Bloom and Weiland, 2014; Bitler et al., 2014; Ding et al., 2016; Walters, 2015; 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 pre-test 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. In addition, Feller et al. (2016) and others argue that that the setting in which a child would otherwise receive care is an important source of impact variation, although this is not directly observable.

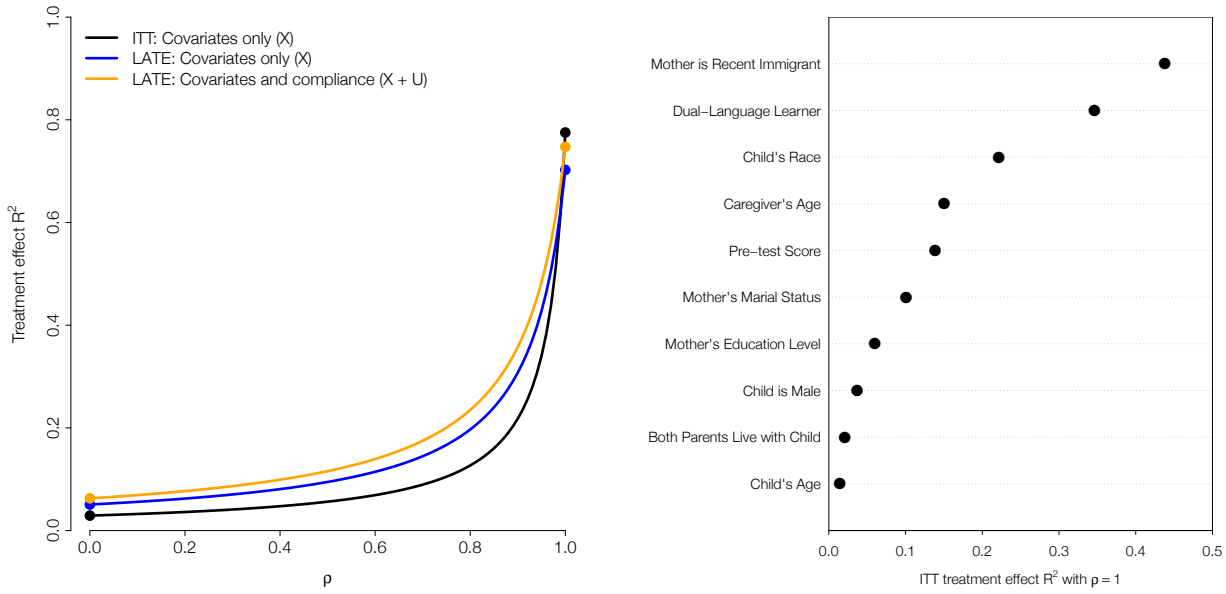
We now use the methods outlined above to assess treatment effect variation in HSIS. The original study included $n = 4,400$ total children, with $n_1 = 2,644$ in the treatment group and $n_0 = 1,796$ in the control group. Following earlier analyses (Ding et al., 2016) and to simplify exposition, we restrict our attention to a complete-case subset of the HSIS, with $n_1 = 2,238$ in the treatment group and $n_0 = 1,348$ 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., re-coding race), the covariate matrix has 17 columns. See Figure 3b 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, $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, and the corresponding model-assisted version of the RI estimator discussed in the Supplementary Material, $\widehat{\boldsymbol{\beta}}^{w}_{\mathrm{RI}}$. 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 $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ estimator and model-assisted $\widehat{\boldsymbol{\beta}}^{w}_{\mathrm{RI}}$ 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 $\widehat{\boldsymbol{\beta}}_{\mathrm{RI}}$ to have the lowest power, it is instructive that the p-value for $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is substantially smaller than the p-value for the



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

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

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 in order 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 3a 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 3b 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 et al. (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 et al. (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 percent 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 3a 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 3a. Taken together, these estimates suggest that there is indeed important treatment effect variation that is neither captured by pre-treatment covariates nor by noncompliance, consistent with previous results in Ding et al. (2016).

8 Conclusion

In this paper, 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 et al., 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, argue 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, i.e., 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 non-randomized 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.

References

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

Administration for Children and Families. 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, 2015.

J. D. Angrist and J. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2008.

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

J. D. Angrist, P. A. Pathak, and C. R. Walters. Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, 5(4):1–27, 2013.

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42:850–871, 2014.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

A. Berrington de González and D. R. Cox. Interpretation of interaction: A review. *The Annals of Applied Statistics*, 1:371–385, 2007.

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

A. S. Blinder. Wage discrimination: reduced form and structural estimates. *Journal of Human resources*, 8:436–455, 1973.

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

J. Bound, D. A. Jaeger, and R. M. Baker. 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, 1995.

W. G. Cochran. *Sampling Techniques*. New York: John Wiley & Sons, 3rd edition, 1977.

D. R. Cox. Interaction (with discussion). *International Statistical Review*, 52:1–24, 1984.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Nonparametric tests for treatment effect heterogeneity. *Review of Economics and Statistics*, 90:389–405, 2008.

J. Currie and D. Thomas. Does Head Start make a difference? *American Economic Review*, 85(3): 341–364, 1995.

P. Ding. A paradox from randomization-based causal inference. *arXiv preprint arXiv:1402.0142*, 2014.

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

H. Djebbari and J. Smith. Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, 145: 64–80, 2008.

A. Feller and A. Gelman. Hierarchical models for causal effects. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, 2015.

A. Feller, T. Grindal, L. Miratrix, and L. C. Page. Compared to what? variation in the impacts of early childhood education by alternative care type. *The Annals of Applied Statistics*, 10(3): 1245–1285, 2016.

R. A. Fisher. *The Design of Experiments.* Edinburgh: Oliver & Boyd, 1st edition, 1935.

C. B. Fogarty. Regression assisted inference for the average treatment effect in paired experiments. *arXiv preprint arXiv:1612.05179*, 2016.

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

M. Fréchet. Sur les tableaux de corrélation dont les marges son données. *Annals Universite de Lyon, Sect. A. Ser. 3*, 14:53–77, 1951.

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

J. Hájek. Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5:361–74, 1960.

J. J. Heckman, J. Smith, and N. Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64:487–535, 1997.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20:217–240, 2011.

W. Hoeffding. Masstabinvariante korrelationsmasse für diskontinuierliche verteilungen. *Arkiv fr matematischen Wirtschaften und Sozialforschung*, 7:49–70, 1941.

Y. Huang, P. B. Gilbert, and H. Janes. Assessing treatment-selection markers using a potential outcomes framework. *Biometrics*, 68:687–696, 2012.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233, 1967.

P. J. Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1:799–821, 1973.

K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7:443–470, 2013.

G. Imbens. Instrumental variables: An econometrician's perspective (with discussion). *Statistical Science*, 29:323–358, 2014.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, and in the Social and Biomedical Sciences*. New York: Cambridge University Press, 2015.

O. Kempthorne. *The Design and Analysis of Experiments*. New York: Wiley, 1952.

E. L. Lehmann. *Elements of Large-Sample Theory*. New York: Springer, 1998.

X. Li and P. Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, page in press, 2017.

W. Lin. Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *The Annals of Applied Statistics*, 7:295–318, 2013.

R. A. Matsouaka, J. Li, and T. Cai. Evaluating marker-guided treatment selection strategies. *Biometrics*, 70:489–499, 2014.

J. A. Middleton and P. M. Aronow. Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy*, 6:39–75, 2015.

R. B. Nelsen. *An Introduction to Copulas*. New York: Springer, 2nd edition, 2007.

J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5:465–472, 1923.

R. Oaxaca. Male-female wage differentials in urban labor markets. *International Economic Review*, 14:693–709, 1973.

M. Puma, S. Bell, R. Cook, C. Heid, G. Shapiro, P. Broene, F. Jenkins, P. Fletcher, L. Quinn, J. Friedman, et al. Head start impact study: Final report. Technical report, Department of Health and Human Services, Administration for Children and Families, Washington DC, 2010.

S. W. Raudenbush and H. S. Bloom. Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, 36(4):475–499, 2015.

P. R. Rosenbaum. Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects. *Biometrics*, 55:560–564, 1999.

P. R. Rosenbaum. *Observational Studies*. New York: Springer, 2nd edition, 2002.

P. R. Rosenbaum. Confidence intervals for uncommon but dramatic responses to treatment. *Biometrics*, 63:1164–1171, 2007.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

D. B. Rubin. Comment on "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, 75:591–593, 1980.

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

D. O. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65:557–586, 1997.

E. A. Stuart, S. R. Cole, C. P. Bradshaw, and P. J. Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

C. R. Walters. Inputs in the production of early childhood human capital: Evidence from head start. *American Economic Journal: Applied Economics*, 7(4):76–102, 2015.

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

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, pages 817–838, 1980.

A. Zhao, P. Ding, R. Mukerjee, and T. Dasgupta. Randomization-based causal inference from split-plot designs. *Annals of Statistics*, page in press, 2017.