JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2021.1978468

Taylor & Francis
Taylor & Francis Group

Noncompliance and Instrumental Variables for 2^K Factorial Experiments

Matthew Blackwell[a] and Nicole E. Pashley[b]

[a]Department of Government and Institute for Quantitative Social Science, Harvard University, Cambridge, MA; [b]Department of Statistics, Rutgers University, Piscataway, NJ

ABSTRACT

Factorial experiments are widely used to assess the marginal, joint, and interactive effects of multiple concurrent factors. While a robust literature covers the design and analysis of these experiments, there is less work on how to handle treatment noncompliance in this setting. To fill this gap, we introduce a new methodology that uses the potential outcomes framework for analyzing 2^K factorial experiments with noncompliance on any number of factors. This framework builds on and extends the literature on both instrumental variables and factorial experiments in several ways. First, we define novel, complier-specific quantities of interest for this setting and show how to generalize key instrumental variables assumptions. Second, we show how partial compliance across factors gives researchers a choice over different types of compliers to target in estimation. Third, we show how to conduct inference for these new estimands from both the finite-population and superpopulation asymptotic perspectives. Finally, we illustrate these techniques by applying them to a field experiment on the effectiveness of different forms of get-out-the-vote canvassing. New easy-to-use, open-source software implements the methodology. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2020
Accepted August 2021

KEYWORDS
Analysis of designed experiments; Causal inference; Factorial experiments, Instrumental variables; Noncompliance

1. Introduction

Researchers across the social and biomedical sciences often rely on factorial experiments to assess the effects of a number of different factors simultaneously. A 2^K factorial experiment randomly assigns units to 2^K possible treatment combinations of K binary factors. These designs have tremendous advantages. First, they allow for the estimation of both the K main effects of each factor and any interactions between the factors. Second, they allow researchers to block certain causal pathways by design and thus provide richer answers to scientific questions. Third, they are also more efficient than experiments that manipulate one factor at a time (Montgomery 2013, chap. 5). Such designs have a long history in statistics (Fisher 1935; Yates 1937) and are often of great scientific and policy relevance. However, only relatively recent literature has begun to address the design and analysis of these experiments under the so-called potential outcomes framework (Hainmueller, Hopkins, and Yamamoto 2014; Dasgupta, Pillai, and Rubin 2015).

A practical consideration with factorial experiments that has received relatively little attention is noncompliance with treatment assignment. This can occur when experimental units self-select into treatment in defiance of their randomized treatment assignment. When this occurs, researchers often switch focus to the intent-to-treat (ITT) effect of treatment assignment. From a scientific and policy viewpoint, however, the primary interest usually remains on the effect of the treatment actually received. In the context of single-factor experiments, researchers can address noncompliance through the use of instrumental variables (IV), which are less frequently used in factorial designs (for exceptions, see Cheng and Small 2006; Blackwell 2017; Schochet 2020). Indeed, the properties of IV estimators in single-factor experiments are well-studied (Angrist, Imbens, and Rubin 1996), but the relevant estimands and estimators have yet to be developed in the factorial case.

We address this problem by introducing a framework for analyzing 2^K factorial experiments with noncompliance on any number of factors. Our contributions are several. First, we generalize the standard instrumental variables framework, including the assumptions and estimands, from the single-factor case to the factorial setting. In particular, we show how to extend key assumptions like the exclusion restriction and monotonicity and how to define novel factorial IV estimands as ratios of intent-to-treat effects of treatment assignment on the outcome and treatment uptake. Unlike the single-factor case, there are several IV estimands in the factorial setting: main effects, two-way interactions, three-way interactions, and so on.

Second, we demonstrate how the multidimensional nature of treatment in factorial experiments complicates the interpretation of these IV estimands. A respondent might comply with their assigned value on one factor but not on another, and the number of possible compliance types grows quickly with K. To address these issues, we invoke an assumption novel to the factorial setting—the "treatment exclusion restriction"—in which the treatment receipt of a factor only depends on the treatment assignment for that factor (Blackwell 2017). Under this and the other IV assumptions, we show that IV estimands

CONTACT Matthew Blackwell *mblackwell@gov.harvard.edu* Harvard University, Government, 1737 Cambridge St, CGIS Knafel Building 305, Cambridge, MA 02143.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

have an interpretation as the average factorial effects of treatment received for the *marginalized compliers*—that is, those respondents who comply with treatment assignment on the active factor(s) for the main effect or interaction of interest, marginalizing over the compliance status of the other factors. One disadvantage of these effects is that the compliance group changes across the different factorial effects, and so we also introduce effects for those that would comply with assignments on all factors, whom we call *perfect compliers*, and develop methods for comparing the different compliance types in terms of their covariate distributions.

Third, to conduct estimation and inference for these IV quantities, we explore two different frameworks: finite-population (also known as finite-sample) inference and superpopulation inference. Following Dasgupta, Pillai, and Rubin (2015) and Kang, Peck, and Keele (2018), our finite-population approach treats the potential outcomes and causal effects of interest as fixed quantities about a finite population. Variation and uncertainty in this approach come only from the random assignment of treatment. We use recent work on finite-population asymptotics to derive a central limit result for our intent-to-treat effects and use this to develop a procedure for generating confidence intervals based on inverting a test involving the intent-to-treat effects (Fieller 1954; Li and Ding 2017; Kang, Peck, and Keele 2018). Superpopulation approaches, on the other hand, assume that the potential outcomes are random draws from an infinite superpopulation, simplifying inference considerably at the price of plausibility.

We then apply our methodology to a get-out-the-vote experiment from New Haven, CT designed to estimate the effects of three treatment factors on voter turnout: door-to-door in person canvassing, phone calls, and mailers (Gerber and Green 2000). While households were randomly assigned to different combinations of voter outreach, many households never received the treatments because they failed to answer the phone or the door. This noncompliance complicates estimation of treatment effects when compliance rates differs across the types of contact. Another empirical application, presented in the supplemental material, uses data from Blattman, Jamison, and Sheridan (2017) to assess the effect of cash transfers and cognitive behavioral therapy on various types of criminal or violent behavior in the short and long term.

The article proceeds as follows. In Section 2, we introduce the setting of factorial experiments with noncompliance and outline our key assumptions, quantities of interest, and estimators. Next, in Section 3, we develop the asymptotic properties of the estimators for the instrumental variable estimands under a finite-population framework and discuss how to apply a technique from the literature on ratio estimators to construct confidence intervals. In Section 4, we describe how to compare different compliance groups in terms of their covariate distributions and present one way to potentially adjust for these differences. We apply all of these techniques to the voter mobilization application in Section 5 and end with concluding thoughts in Section 6. In the Supplemental Materials, we also develop a procedure for Bayesian inference in this context and present simulation evidence for the validity of our confidence interval procedure.

2. Framework

We consider an experiment with K binary factors with levels $\{-1, +1\}$, so that $\mathcal{Z} = \{-1, +1\}^K$ is the set of all possible treatment combinations. For instance, -1 may be the control level and $+1$ the treatment level of a given factor. Thus, there are $L = 2^K$ possible treatment assignments, which we order $\{1, \dots, L\}$ with $\boldsymbol{z}_\ell = \{z_{\ell 1}, \dots, z_{\ell K}\}$ being the levels of each factor for treatment combination ℓ. We define the set of possible treatment uptake vectors $\boldsymbol{d}_\ell$, which have the same values and are ordered in the same manner as $\boldsymbol{z}_\ell$ (i.e., $\boldsymbol{d}_\ell = \boldsymbol{z}_\ell$). Each unit may have a different potential outcome for each treatment assignment and uptake combination, $Y_i(\boldsymbol{d}, \boldsymbol{z})$. This is the value of the outcome that unit i would have if they been assigned $\boldsymbol{z}$ and taken $\boldsymbol{d}$.

Experiments with noncompliance face the problem that treatment uptake may differ from treatment assignment, and so treatment uptake will have potential outcomes as well. Let $\boldsymbol{D}_i(\boldsymbol{z}) \in \mathcal{Z}$ be the vector of treatment uptake on each factor if unit i was assigned to treatment combination $\boldsymbol{z}$. If $\boldsymbol{D}_i(\boldsymbol{z}) = \boldsymbol{z}$ for all i and $\boldsymbol{z}$, then there is full compliance and inference can be conducted as usual. We focus on the case where $\boldsymbol{D}_i(\boldsymbol{z}) \neq \boldsymbol{z}$ for some i and $\boldsymbol{z} \in \mathcal{Z}$ and define the vector of potential outcomes indicators for each treatment uptake combination as

$$\boldsymbol{R}_i(\boldsymbol{z}) = \{\mathbb{I}(\boldsymbol{D}_i(\boldsymbol{z}) = \boldsymbol{d}_1), \dots, \mathbb{I}(\boldsymbol{D}_i(\boldsymbol{z}) = \boldsymbol{d}_L)\}^\top .$$

Let $\boldsymbol{R}_i(\bullet)$ be the $2^K \times 2^K$ matrix with ℓth row $\boldsymbol{R}_i(\boldsymbol{z}_\ell)^\top$. For the intent-to-treat analyses, we will often work with the potential outcomes just setting the treatment assignment, $Y_i(\boldsymbol{z}) \equiv Y_i(\boldsymbol{z}, \boldsymbol{D}_i(\boldsymbol{z}))$, and we collect the L potential outcomes for unit i into the vector $\boldsymbol{Y}_i(\bullet) = \{Y_i(\boldsymbol{z}_1), \dots, Y_i(\boldsymbol{z}_L)\}^\top$.

Let $W_{i\ell} = 1$ if $\boldsymbol{Z}_i = \boldsymbol{z}_\ell$ and 0 otherwise and $\boldsymbol{W}_i = \{W_{i1}, \dots, W_{iL}\}$ be the vector of indicators for all treatment combinations. We assume a completely randomized design. In particular, let $\boldsymbol{W} = (\boldsymbol{W}_1, \dots, \boldsymbol{W}_N)$ be the length LN vector of assignment indicators for all units and $\mathcal{F} = \{\boldsymbol{Y}_i(\bullet), \boldsymbol{R}_i(\bullet), i = 1, \dots, N\}$. Consider a completely randomized design with $N_\ell = \sum_{i=1}^N W_{i\ell}$ units assigned to treatment $\boldsymbol{z}_\ell$, with $\sum_{\ell=1}^L N_\ell = N$, defined formally below.

Assumption 1 (General completely randomized design).

$$\mathbb{P}(\boldsymbol{W} | \mathcal{F}) = \begin{cases} \left(\dfrac{N!}{\prod_{\ell=1}^L N_\ell !} \right)^{-1} & \text{if } \sum_{i=1}^N W_{i\ell} = N_\ell \text{ for all} \\ & \ell = 1, \dots, L \\ 0 & \text{otherwise} \end{cases}$$

Under this design, we have $\mathbb{E}\{W_{i\ell} | \mathcal{F}\} = N_\ell / N$ for all $\boldsymbol{z}_\ell$, where the expectation here is over the randomization distribution. We connect the potential outcomes to the observed outcomes through a consistency assumption, $Y_i^{\text{obs}} = \sum_{\ell=1}^L W_{i\ell} Y_i(\boldsymbol{z}_\ell)$, $\boldsymbol{D}_i^{\text{obs}} = \sum_{\ell=1}^L W_{i\ell} \boldsymbol{D}_i(\boldsymbol{z}_\ell)$, and $\boldsymbol{R}_i^{\text{obs}} = \sum_{\ell=1}^L W_{i\ell} \boldsymbol{R}_i(\boldsymbol{z}_\ell)$, which implicitly assumes the stable unit treatment value assumption (Rubin 1980).

When there is noncompliance with treatment assignment, randomization is not sufficient to identify the causal effect of treatment uptake. Several ways of addressing noncompliance have been proposed in the literature, all of which make additional assumptions beyond randomization. We follow one strain of the literature, which started with Angrist, Imbens, and Rubin (1996), and focus on two types of assumptions: monotonicity and exclusion restrictions. We generalize these standard instrumental variables assumptions to the factorial context.

Monotonicity is a restriction on the direction of the effect of treatment assignment on treatment uptake. Let $\boldsymbol{z}^+$ be a K-vector of all $+1$ and $\boldsymbol{z}^-$ be a K-vector of all -1, with z_k^+ and z_k^- being representative kth entries. Furthermore, let $\boldsymbol{z}_{-k}$ be the vector $\boldsymbol{z}$ with the kth entry omitted and abuse notation to let $\boldsymbol{z} = (\boldsymbol{z}_{-k}, z_k)$. Let $D_{ik}(\boldsymbol{z})$ be the treatment uptake of unit i for factor k when assigned to $\boldsymbol{z}$.

Assumption 2 (Monotonicity). $D_{ik}(\boldsymbol{z}_{-k}, z_k^+) \geq D_{ik}(\boldsymbol{z}_{-k}, z_k^-)$ for all $k \in \{1, \ldots, K\}$ and $\boldsymbol{z}_{-k}$.

This assumption states that there are no defiers: individuals who would have treatment uptake of -1 for factor k if assigned to $+1$ of factor k and treatment uptake of $+1$ for factor k if assigned to -1 of factor k, holding the assignment of the other factors constant.

A standard approach in the instrumental variables literature is to assume that treatment assignment has no direct effect on the outcome, except through treatment receipt (Robins 1989; Angrist, Imbens, and Rubin 1996). This assumption is typically called the *exclusion restriction*, and it has a natural generalization in the factorial setting. To distinguish it from a separate exclusion restriction we define below, we call this the outcome exclusion restriction.

Assumption 3 (Outcome exclusion restriction). For all $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$, $Y_i(\boldsymbol{z}, \boldsymbol{d}) = Y_i(\boldsymbol{z}', \boldsymbol{d})$.

This assumption is substantive and cannot be met simply by experimental design. Finally, the factorial setting requires a novel assumption for identification of certain effects. First proposed in Blackwell (2017) for the 2 × 2 factorial design, the *treatment exclusion restriction* states that treatment uptake on factor k only depends on the treatment assignment for factor k, not other factors.

Assumption 4 (Treatment exclusion restriction). For all $\boldsymbol{z} \in \mathcal{Z}$, $D_{ik}(\boldsymbol{z}) = D_{ik}(z_k)$ where z_k is the kth entry of $\boldsymbol{z}$.

This assumption restricts compliance to be factor-specific, and prevents any factor from affecting the uptake on another factor. Furthermore, it rules out interactive effects of treatment assignment on treatment uptake in the sense that it assumes no units that, say, comply on factor 1 when $z_2 = +1$ but not when $z_2 = -1$. The treatment exclusion restriction is a substantive assumption that restricts the first-stage relationship between treatment assignment and treatment uptake. In the context of the voter mobilization experiment, for instance, this would be violated if being assigned to receive door-to-door contact caused some respondents to pick up for a phone contact attempt when they otherwise would not. While treatment exclusion is not directly testable, some of its implications are observable. For instance, it would rule out any effect of Z_{i1} on D_{i2} or any interaction between Z_{i1} and Z_{i2} on D_{i2}. Thus, one falsification test for this assumption is to check these various effects in the assignment-uptake relationship, which we do in our empirical example below. We discuss some implications for weakening this assumption in the following section and outline further weaker assumptions of interest in the Discussion.

2.1. Estimands

We begin by describing a set of standard linear factorial effects in the finite-population framework and then extend them to the superpopulation viewpoint below. These effects reflect differences between one half of the potential outcomes for a particular outcome versus the other. We can define these effects through the use of an L-dimensional vector $\boldsymbol{g}$ that has one half of its entries at 1 and the other half at -1 as in Dasgupta, Pillai, and Rubin (2015). There are $L - 1$ such vectors and the same number of factorial effects. We can order these vectors such that the first K represent the main effects of the K factors, so that $\boldsymbol{g}_1$ corresponds to the main effect of factor 1, $\boldsymbol{g}_2$ corresponds to the main effect of factor 2, and so on. The next $\binom{K}{2}$ vectors will correspond to all two-factor interactions, and the following $\binom{K}{3}$ vectors will correspond to all three-factor interactions, and so on. This continues until $\boldsymbol{g}_{L-1}$ which corresponds to the K-way interaction between all factors. For main effects, $\boldsymbol{g}_j$ is a vector giving the level of factor j for each of the L treatment combinations. Interaction vectors are then created as element-wise products of these main effect vectors. Note that these vectors are mutually orthogonal.

With these vectors, we define individual-level intent-to-treat factorial effects for the outcome as

$$\tau_{ij} = 2^{-(K-1)} \boldsymbol{g}_j^\top \boldsymbol{Y}_i(\bullet) = 2^{-(K-1)} \sum_{\ell=1}^{L} g_{j\ell} Y_i(\boldsymbol{z}_\ell)$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, L-1$, where $g_{j\ell}$ is the ℓth entry of the $\boldsymbol{g}_j$ vector. Here, τ_{ij} is the jth factorial effect of treatment assignment on the outcome for individual i. For main effects, this is the effect of assignment to factor j, averaging over all possible assignments to other factors. For example, when $K = 2$, we have $\boldsymbol{g}_1 = (+1, -1, +1, -1)$, so that

$$\frac{1}{2} \boldsymbol{g}_1^\top \boldsymbol{Y}_i(\bullet) = \frac{1}{2} \underbrace{\{Y_i(+1,+1) - Y_i(-1,+1)\}}_{\text{effect of factor 1 when factor 2 is } +1} + \frac{1}{2} \underbrace{\{Y_i(+1,-1) - Y_i(-1,-1)\}}_{\text{effect of factor 1 when factor 2 is } -1}.$$

Writing the finite-population averages of the potential outcomes as $\overline{\boldsymbol{Y}}(\bullet) = N^{-1} \sum_{i=1}^{N} \boldsymbol{Y}_i(\bullet)$, the finite-population intent-to-treat average factorial effects on the outcome are

$$\overline{\tau}_j = \frac{1}{N} \sum_{i=1}^{N} \tau_{ij} = 2^{-(K-1)} \boldsymbol{g}_j^\top \overline{\boldsymbol{Y}}(\bullet).$$

These effects marginalize over treatment assignment on the other factors, weighting each possible assignment equally. While this is standard in the factorial design literature, recent work on a specific type of factorial designs—conjoint experiments—has dealt with a more general estimand that allows for researcher-specified distributions for the assignments (Hainmueller, Hopkins, and Yamamoto 2014; de la Cuesta, Egami, and Imai 2021). In the supplemental material, we discuss the straightforward extension of the present approach to those more general estimands. Finally, Egami and Imai (2019) proposed alternative quantities of interest for interactions in factorial experiments, but those average marginal interaction effects are more appropriate with factors with more than two levels.

These intent-to-treat factorial effects will not equal the true effect of treatment uptake when some units do not comply with the factors in the factorial effect. To correct this problem, the instrumental variables literature will often define the estimand of interest as the ratio of the intent-to-treat effects on the outcome and on treatment uptake (Wald 1940). In the factorial setting, however, the definition of treatment uptake depends on the factorial effect of interest. For example, for the main effect of the first factor, we want the ITT for treatment uptake on the first factor, whereas for the interaction between the first and second factor, we want the ITT on the *interaction* between D_{i1} and D_{i2}. More generally, let $\mathcal{K}(j)$ be the set of indices of the "active" factors for factorial effect j. That is, $\mathcal{K}(j)$ are the set of factors for which $\boldsymbol{g}_j$ is estimating the main or interaction effects. For the main effects, $j = 1, \dots, K$, this is just $\mathcal{K}(j) = \{j\}$, but for interactions, we have for example, $\mathcal{K}(K+1) = \{1, 2\}$, and so on. Define the following potential outcome of treatment uptake interaction corresponding to the jth factorial effect:

$$\widetilde{D}_{ij}(\boldsymbol{z}) = \prod_{k \in \mathcal{K}(j)} D_{ik}(\boldsymbol{z}).$$

Again, for $j \leq K$, we have $\widetilde{D}_{ij}(\boldsymbol{z}) = D_{ij}(\boldsymbol{z})$. We can collect these into a vector of potential outcomes for each treatment assignment vector $\widetilde{\boldsymbol{D}}_{ij}(\bullet) = \{\widetilde{D}_{ij}(\boldsymbol{z}_1), \dots, \widetilde{D}_{ij}(\boldsymbol{z}_L)\}^\top$. Further, as we show in supplemental material A, we can write these as a function of the $\boldsymbol{g}$ vectors to obtain $\widetilde{D}_{ij}(\boldsymbol{z}) = \boldsymbol{g}_j^\top \boldsymbol{R}_i(\boldsymbol{z})$ since, by construction, $\boldsymbol{g}_j$ is equal to the product of the active factors for each of the possible vectors of treatment uptake and $\boldsymbol{R}_i(\boldsymbol{z})$ indicates which of these assignment vectors is selected for unit i based on their compliance type. Furthermore, this implies $\widetilde{\boldsymbol{D}}_{ij}(\bullet) = \boldsymbol{R}_i(\bullet)\boldsymbol{g}_j$. The individual-level ITT of treatment assignment on treatment uptake for the jth factorial effect is thus

$$\delta_{ij} = 2^{-K} \boldsymbol{g}_j^\top \widetilde{\boldsymbol{D}}_{ij}(\bullet) = 2^{-K} \boldsymbol{g}_j^\top \boldsymbol{R}_i(\bullet)\boldsymbol{g}_j,$$

with $\overline{\delta}_j = N^{-1} \sum_{i=1}^N \delta_{ij}$. For example, in the two-factor case, we have

$$\delta_{i3} = \frac{1}{4}\{D_{i1}(+1,+1)D_{i2}(+1,+1) - D_{i1}(-1,+1)D_{i2}(-1,+1)\}$$
$$-\frac{1}{4}\{D_{i1}(+1,-1)D_{i2}(+1,-1) - D_{i1}(-1,-1)D_{i2}(-1,-1)\},$$

so that δ_{i3} is the (scaled) interactive effect of treatment assignment on the multiplicative interaction between the two treatment uptakes. We can also write this estimand as a linear function of the potential outcomes for each assignment, $\delta_{ij} = \sum_{\ell=1}^L 2^{-K} g_{j\ell} \boldsymbol{g}_j^\top \boldsymbol{R}_i(\boldsymbol{z}_\ell)$, where the equality comes from $\widetilde{D}_{ij}(\boldsymbol{z}) = \boldsymbol{g}_j^\top \boldsymbol{R}_i(\boldsymbol{z})$ and $g_{j\ell}$ is the ℓth entry of $\boldsymbol{g}_j$.

We can now define the jth IV factorial effect as

$$\overline{\phi}_j = \frac{\overline{\tau}_j}{\overline{\delta}_j}.$$

We assume that $\overline{\delta}_j > 0$, which under treatment exclusion means that there are *some* compliers for the factors involved in the jth effect. Without further assumptions, $\overline{\phi}_j$ is just the ratio of two intent-to-treat factorial effects. We are able to gain an even more substantive interpretation under various exclusion restrictions on the outcome and the treatment uptake, as described in the next section.

2.2. Interpretation of the Estimands Under IV Assumptions

Under the IV assumptions, the various effects defined above have specific interpretations in terms of principal strata, otherwise known as compliance types. Under treatment exclusion and monotonicity, each unit can be categorized into one of 3^K types based on how treatment uptake depends on treatment assignment. Note that without the treatment exclusion restriction we would have many more compliance types, as a unit's compliance to a given factor could depend upon the 2^{K-1} possible assignments to the other factors. Thus, the treatment exclusion assumption essentially makes solutions based on compliance strata more tractable. Let $\boldsymbol{T}_i \in \mathcal{T}_K = \{c, a, n\}^K$ be the K-length vector of compliance type for unit i on all K factors. Here, the compliance types of each factor are complier (c), always-taker (a), and never-taker (n), defined as follows:

$$T_{ik} = \begin{cases} c & \text{if } D_{ik}(+1) = +1, D_{ik}(-1) = -1 \\ a & \text{if } D_{ik}(+1) = +1, D_{ik}(-1) = +1 \\ n & \text{if } D_{ik}(+1) = -1, D_{ik}(-1) = -1. \end{cases}$$

Our estimands relate to these quantities in two key ways. First, under treatment exclusion and monotonicity, for any factorial effect, we have $\widetilde{\boldsymbol{D}}_{ij}(\bullet) = \boldsymbol{g}_j$ when $T_{ik} = c$ for all $k \in \mathcal{K}(j)$ and otherwise $\widetilde{\boldsymbol{D}}_{ij}(\bullet)$ is a vector that is orthogonal to $\boldsymbol{g}_j$. We define $C_{ij} = \prod_{k \in \mathcal{K}(j)} \mathbb{I}(T_{ik} = c)$, an indicator for being a complier on all the active factors for effect j. Then for all j, we have $\delta_{ij} = C_{ij}$ and $\overline{\delta}_j = N^{-1} \sum_{i=1}^N C_{ij}$. We provide a more formal proof of this result in supplemental material A. In other words, the ITTs for treatment uptake measure compliance with the active factors for a particular factorial effect.

Second, under monotonicity and the treatment and outcome exclusion restrictions, the jth outcome ITT, τ_{ij}, is 0 for all units who do not comply on all the active factors in effect j, allowing us to relate these effects to the conditional effect among compliers. Let $N_j^c = \sum_{i=1}^N C_{ij}$. Noting that $\overline{\delta}_j = N_j^c / N$, we have the following:

$$\overline{\tau}_j = \frac{1}{N} \sum_{i=1}^N C_{ij} \tau_{ij} = \frac{\sum_{i=1}^N C_{ij} \tau_{ij}}{N_j^c} \times \overline{\delta}_j.$$

Combining these two facts, the ratio of the ITT effects under the IV assumptions (Assumptions 2, 3, and 4) is

$$\overline{\phi}_j = \frac{1}{N_j^c} \sum_{i=1}^N C_{ij} \tau_{ij},$$

which we refer to as the jth marginalized-complier average factorial effect (MCAFE). Because these effects condition on compliance for the active factors, we can interpret this as the average of the jth factorial effect of treatment uptake of factors in $\mathcal{K}(j)$ on the outcome among those units who comply with those active factors, marginalizing over the treatment assignments on other factors. For a main effect, for instance, we show in supplemental material A that

$$\overline{\phi}_j = \frac{1}{2^{K-1}} \sum_{\boldsymbol{z}_{-j} \in \mathcal{Z}_{-j}} \left(\frac{1}{N_j^c} \sum_{i=1}^{N} C_{ij} \left\{ Y_i(d_j = +1, \boldsymbol{z}_{-j}) - Y_i(d_j = -1, \boldsymbol{z}_{-j}) \right\} \right),$$

where $\boldsymbol{z}_{-j}$ is the assignment vector $\boldsymbol{z}$ less the entry for factor j and $\mathcal{Z}_{-j}$ is the associated set of possible such assignments. Here, we slightly abuse notation to emphasize that it is truly treatment uptake, and not just assignment for factor j. This interpretation, while straightforward to derive, is slightly odd because it combines the effects of treatment uptake for some factors and treatment assignment for others.

How can we interpret the MCAFEs in terms of the factorial effects of treatment uptake rather than a mix of treatment uptake and assignment? We can invoke the exclusion restrictions to write the main effect MCAFEs, for instance, as

$$\begin{aligned}
\overline{\phi}_j &= \frac{1}{2^{K-1}} \sum_{\boldsymbol{z}_{-j}} \left(\frac{1}{N_j^c} \sum_{i=1}^{N} C_{ij} \left\{ Y_i(d_j = +1, \boldsymbol{D}_{i,-j}(\boldsymbol{z}_{-j})) \right. \right. \\
&\qquad \left. \left. - Y_i(d_j = -1, \boldsymbol{D}_{i,-j}(\boldsymbol{z}_{-j})) \right\} \right), \\
&= \sum_{\boldsymbol{d}_{-j}} \left(\frac{1}{N_j^c} \sum_{i=1}^{N} \omega_{ij}(\boldsymbol{d}_{-j}) C_{ij} \left\{ Y_i(d_j = +1, \boldsymbol{d}_{-j}) \right. \right. \\
&\qquad \left. \left. - Y_i(d_j = -1, \boldsymbol{d}_{-j}) \right\} \right),
\end{aligned}$$

where

$$\omega_{ij}(\boldsymbol{d}_{-j}) = \frac{1}{2^{K-1}} \sum_{\boldsymbol{z}_{-j}} \mathbb{I}\left\{ \boldsymbol{D}_{i,-j}(\boldsymbol{z}_{-j}) = \boldsymbol{d}_{-j} \right\},$$

$$\sum_{\boldsymbol{d}_{-j}} \omega_{ij}(\boldsymbol{d}_{-j}) = 1.$$

We again commit slight abuse of notation to convey the meaning in terms of treatment uptake rather than assignment. Thus, we can see that the MCAFE for the main effect of factor j is an average of complier factorial effects for treatment uptake with each individual having different weights for marginalizing over the uptake profiles. These weights depend on the unit's compliance type on the other factors. Interpretations of the higher-order MCAFEs are similar, albeit more complicated.

Of course, treatment exclusion is a strong assumption that may not hold in practice, so it is helpful to understand how we can interpret these IV estimands under weaker assumptions. In supplemental materials C, we show the IV estimands retain a similar, though much more complicated, interpretation as a weighted average of effects under a weaker version of the treatment exclusion assumption. Unfortunately, the interpretation of interactions under this weaker assumption is much less clear, which highlights how identifying interactive effects of treatment uptake requires restrictions on interactions of treatment assignment on treatment uptake.

2.3. Disadvantages of MCAFEs

One important disadvantage of marginalized-complier effects is that the conditioning set changes depending on the factorial effect under study. This makes, for instance, the main effect of factor 1 and the interaction effect of factor 1 and 2 difficult to compare. The first MCAFE will only condition on compliers for factor 1 and average over compliance groups for factor 2, while the latter will focus on compliers for factor 1 and factor 2. If the complier groups differ significantly between factorial effects, it is impossible to tell if differences between factorial effects are due to true differences in average effects or simply manifestations of heterogeneous treatment effects across compliance types. This is especially problematic for factorial experiments, where much of the value comes from comparing effects both within orders (the effect of factor 1 vs the effect of factor 2) and between them (main effects vs interactions).

2.4. Perfect Complier Effects

One way to avoid the disadvantages of the effects for marginal compliers is to estimate effects for those units that would comply with all factors—whom we call *perfect compliers*. The main advantage of this approach is that every factorial effect is well-defined for the perfect compliers. Thus, comparing different factorial effects in this subset will not be driven by changes in the compliance groups as with marginal compliers. One of the main disadvantages of working with perfect compliers is that, by definition, there are fewer of them than marginal compliers, leading to greater uncertainty in our inferences. Another disadvantage is that the IV estimands for perfect compliers are not simply a ratio of ITT effects on the outcome to ITT effects on treatment uptake. At first glance, it may appear that focusing on perfect compliers simplifies our task since we have reduced our very complicated compliance problem to a single binary compliance problem. Unfortunately, while there is only one way to be a perfect complier, there are still many ways to be a non-perfect-complier and so isolating just the effects for perfect compliers requires more care than simply using existing 2^K factorial methods.

To start, we can (given all potential outcomes) identify the perfect compliers by applying the K-way interaction to any vector of potential outcomes for specific treatment uptake vectors under the IV assumptions discussed earlier. Specifically, let $P_i = \prod_{k=1}^{K} \mathbb{I}(T_{ik} = c)$ be an indicator for being a perfect complier. From the above discussion, the marginalized compliers for the K-way interaction will be the perfect compliers, so $\delta_{i,L-1} = P_i$. In order to identify the potential outcomes among the perfect compliers, we must modify the ITT for the outcome. Let $\boldsymbol{H}_i(\boldsymbol{z}) = Y_i(\boldsymbol{z})\boldsymbol{R}_i(\boldsymbol{z})$ so that

$$\boldsymbol{H}_i(\boldsymbol{z}) = \{Y_i(\boldsymbol{z})\mathbb{I}(\boldsymbol{D}_i(\boldsymbol{z}) = \boldsymbol{d}_1), \ldots, Y_i(\boldsymbol{z})\mathbb{I}(\boldsymbol{D}_i(\boldsymbol{z}) = \boldsymbol{d}_L)\}^\top,$$

and let $\boldsymbol{H}_i(\bullet)$ be the $L \times L$ matrix with ℓth row $\boldsymbol{H}_i(\boldsymbol{z}_\ell)^\top$. We show in supplemental material A that, under Assumptions 2– 4, we have

$$\boldsymbol{g}_{L-1} \circ \boldsymbol{H}_i(\bullet)^\top \boldsymbol{g}_{L-1} = \boldsymbol{Y}_i^d(\bullet) P_i,$$

where $\boldsymbol{Y}_i^d(\bullet) = \{Y_i(\boldsymbol{d}_1), \ldots, Y_i(\boldsymbol{d}_L)\}$ are the vector of potential outcomes as functions of treatment uptake alone. Thus, we can write the jth ITT for unit i, if unit i is a perfect complier, as

$$\tau_{ij,p} = 2^{-(K-1)} \left(\boldsymbol{g}_j \circ \boldsymbol{g}_{L-1}\right)^\top \boldsymbol{H}_i(\bullet)^\top \boldsymbol{g}_{L-1} = P_i \tau_{ij}.$$

In order to isolate the effects for perfect compliers, $\tau_{ij,p}$ involves a complicated interaction effect of treatment assignment on products of the outcome and treatment uptake, rather than sharing the form of the typical factorial effects on Y_i. As with τ_{ij} and δ_{ij}, we can write this quantity as a linear function of the potential outcomes for each assignment,

$$\tau_{ij,p} = \sum_{\ell=1}^{L} g_{L-1,\ell} \left(\boldsymbol{g}_j \circ \boldsymbol{g}_{L-1}\right)^\top \boldsymbol{H}_i(\boldsymbol{z}_\ell),$$

again where $g_{L-1,\ell}$ is the ℓth entry in the $\boldsymbol{g}_{L-1}$ vector. Let $\overline{\boldsymbol{H}}(\bullet) = N^{-1} \sum_{i=1}^N \boldsymbol{H}_i(\bullet)$ the be population average of $\boldsymbol{H}_i(\bullet)$. Then we can define the population effects as

$$\begin{aligned} \overline{\tau}_{j,p} &= 2^{-(K-1)} \left(\boldsymbol{g}_j \circ \boldsymbol{g}_{L-1}\right)^\top \overline{\boldsymbol{H}}(\bullet)^\top \boldsymbol{g}_{L-1} \\ &= \tfrac{1}{N} \textstyle\sum_{i=1}^N P_i \tau_{ij} = \left(\tfrac{1}{N_p} \textstyle\sum_{i=1}^N P_i \tau_{ij}\right) \tfrac{N_p}{N}, \end{aligned}$$

where N_p is the number of perfect compliers in the finite population. Noting from our earlier discussion that $\overline{\delta}_{L-1} = N_p/N$, we can define

$$\overline{\gamma}_j = \frac{\overline{\tau}_{j,p}}{\overline{\delta}_{L-1}} = \frac{1}{N_p} \sum_{i=1}^N P_i \tau_{ij}$$

The $\overline{\gamma}_j$ represents the jth average factorial effect among the perfect compliers, which we refer to as the jth perfect complier average factorial effect (PCAFE). For both the PCAFE and the MCAFE, we cannot identify who is and is not a complier, but in Section 4.1 we show how to estimate covariate profiles of these groups to aid in the interpretability of these effects.

2.5. Superpopulation Estimands

We now take an alternative point of view—that the sample of units is actually a draw from an infinite superpopulation. Now, the potential outcomes are themselves random variables and not fixed quantities as in the finite-population point of view. Under treatment exclusion in particular, we define the probability of a particular compliance type, $\boldsymbol{t} \in \mathcal{T}_K$ as $\rho_{\boldsymbol{t}} = \mathbb{P}(\boldsymbol{T}_i = \boldsymbol{t})$. We can relate the finite-population quantities $\overline{\delta}_k$ to these values by considering the limit of a series of growing finite populations with units sampled from a larger fixed population. For example, for any main effect, we have

$$\lim_{N\to\infty} \overline{\delta}_k = \sum_{\boldsymbol{t}:t_k=c} \rho_{\boldsymbol{t}} = \mathbb{P}(C_{ik} = 1).$$

Let $\mathbb{E}[\cdot]$ be the expectation operator that averages over both randomization and sampling from the superpopulation. Then, we can define the superpopulation version of the marginalized-complier average factorial effect as

$$\overline{\phi}_j^{\text{sp}} = \mathbb{E}[\tau_{ij} | C_{ij} = 1] = \lim_{N\to\infty} \overline{\phi}_j.$$

We can define a similar superpopulation version of the perfect complier average factorial effect as

$$\overline{\gamma}_j^{\text{sp}} = \mathbb{E}[\tau_{ij} | P_i = 1] = \lim_{N\to\infty} \overline{\gamma}_j.$$

Finally, we can define $\overline{\tau}_j^{\text{sp}}$, $\overline{\tau}_{j,p}^{\text{sp}}$, and $\overline{\delta}_j^{\text{sp}}$ in a similar manner.

2.6. Estimators

We can define the following natural in-sample estimators for the population (of units in the study) or superpopulation potential outcomes:

$$\overline{Y}^{\text{obs}}(\boldsymbol{z}_\ell) = \frac{1}{N_\ell} \sum_{i=1}^N W_{i\ell} Y_i^{\text{obs}}, \qquad \overline{\boldsymbol{R}}^{\text{obs}}(\boldsymbol{z}_\ell) = \frac{1}{N_\ell} \sum_{i=1}^N W_{i\ell} \boldsymbol{R}_i^{\text{obs}},$$

$$\overline{\boldsymbol{H}}^{\text{obs}}(\boldsymbol{z}_\ell) = \frac{1}{N_\ell} \sum_{i=1}^N W_{i\ell} \boldsymbol{H}_i(\boldsymbol{z}_\ell).$$

These lead to the natural estimators for the various ITT effects:

$$\widehat{\tau}_j = \sum_{\ell=1}^L 2^{-(K-1)} g_{j\ell} \overline{Y}^{\text{obs}}(\boldsymbol{z}_\ell), \qquad \widehat{\delta}_j = \sum_{\ell=1}^L 2^{-K} g_{j\ell} \boldsymbol{g}_j^\top \overline{\boldsymbol{R}}^{\text{obs}}(\boldsymbol{z}_\ell),$$

$$\widehat{\tau}_{j,p} = \sum_{\ell=1}^L 2^{-(K-1)} g_{L-1,\ell} \left(\boldsymbol{g}_j \circ \boldsymbol{g}_{L-1}\right)^\top \overline{\boldsymbol{H}}^{\text{obs}}(\boldsymbol{z}_\ell).$$

Under a completely randomized design, we have $\mathbb{E}\left[\overline{Y}^{\text{obs}}(\boldsymbol{z}) | \mathcal{F}\right] = \overline{Y}(\boldsymbol{z})$, which implies that $\widehat{\tau}_j$ is unbiased for $\overline{\tau}_j$ when averaging over the randomization distribution. The same result holds for $\widehat{\delta}_j$ and $\widehat{\tau}_{j,p}$ for $\overline{\delta}_j$ and $\overline{\tau}_{j,p}$, respectively. Importantly, these results do not depend on any of the instrumental variable assumptions and hold by experimental design. Finally, we can define estimators for the MCAFE and the PCAFE as:

$$\widehat{\phi}_j = \widehat{\tau}_j / \widehat{\delta}_j \qquad \widehat{\gamma}_j = \widehat{\tau}_{j,p} / \widehat{\delta}_{L-1}.$$

Each of these estimators has a similar form to the classic Wald estimator: ratios of ITT effects on the outcome to ITT effects on (some function of) treatment uptake.

3. Inference

Inference for instrumental variables estimators has generally followed two broad approaches. First, and more traditionally, one can assume that the data are a random sample from an infinite superpopulation and derive the asymptotic distribution of the various estimators from the central limit theorem and the delta method. This approach has the advantage that the subsamples corresponding to each treatment assignment vector, $\boldsymbol{z}_\ell$, can be thought of as independent random samples from different population distributions, which greatly simplifies derivation of the large-sample distribution of the estimators. This approach considers variation in the estimates both from the randomization of $\boldsymbol{Z}_i$ and the random sampling from the superpopulation. The second approach to inference is to take the finite-population quantities $\overline{\phi}_j$ and $\overline{\gamma}_j$ as the quantities of interest and consider the behavior of the estimators over the distribution of the treatment assignments induced by randomization (Fisher 1935; Imbens and Rosenbaum 2005). This approach has the advantage that it hews closely to the design of the original experiment and is well-defined even when it is difficult to imagine a hypothetical superpopulation. Below, we present results for the finite-population setting and then show how they change when targeting inference to a superpopulation.

Once an asymptotic distribution has been established, there are several ways to construct confidence intervals for the types of ratio estimators we defined above. The standard way to construct confidence intervals for, say, $\widehat{\phi}_j$ would be to use the delta method on the ratio of $\widehat{\tau}_j$ and $\widehat{\delta}_j$ to obtain an estimator of its asymptotic variance, $\widehat{V}_j$. Then, a 95% confidence interval could be obtained from $\widehat{\phi}_j \pm 1.96 \times \widehat{V}_j$. Unfortunately, this approach, which is based on a Taylor expansion, can be a poor approximation when the denominator is close to 0 (in our case, when there are relatively few compliers). An alternative approach, first proposed by Fieller (1954), uses a carefully chosen test statistic and inverts it to construct the confidence intervals. The key to this approach is that the variance of the test statistic under the null can be written as a quadratic function of a null hypothesis of the true effect, allowing the confidence intervals to achieve nominal coverage even when the denominator is close to zero. The tradeoff is that these confidence intervals can have infinite length in some samples. See supplementary material E for simulations exploring the performance of the different confidence interval methods and for MCAFE vs PCAFE estimators.

3.1. Expectation and Variances in the Finite Population

Although we cannot directly calculate the expectations and variances of our ratio estimators in the finite population, we can derive these properties for their numerators and denominators. Let $\boldsymbol{U}_i(\boldsymbol{z}) = \{\boldsymbol{H}_i(\boldsymbol{z}), \boldsymbol{R}_i(\boldsymbol{z})\}^\top$ be the vector of all $2L$ potential outcomes for unit i under treatment assignment $\boldsymbol{z}$ and let $\overline{\boldsymbol{U}}(\boldsymbol{z})$ be the vector of $2L$ finite-population means. Similarly, let $\widehat{\boldsymbol{U}}(\boldsymbol{z})$ be the vector of estimated means based on treatment assignment. All of the ITT quantities of interest defined in previous sections are linear combinations of these potential outcomes.

Combining all of the above estimands, we are interested in $r = 3L - 3$ of these effects; $L-1$ intent-to-treat factorial effects on the outcome, $\overline{\tau}_j$, $L-1$ effects among the perfect compliers, $\overline{\tau}_{j,p}$, and $L-1$ intent-to-treat effects on the treatment uptake indicators, $\overline{\delta}_j$. As in Li and Ding (2017), we can write our vector of estimands using coefficient matrices $\boldsymbol{Q}_\ell \in \mathbb{R}^{r \times 2L}$ so that we have

$$\boldsymbol{\theta}_i = \sum_{\ell=1}^{L} \boldsymbol{Q}_\ell \boldsymbol{U}_i(\boldsymbol{z}_\ell)$$

$$\boldsymbol{\theta}_i = \{\tau_{i1}, \ldots, \tau_{i,L-1}, \tau_{i1,p}, \ldots, \tau_{i,L-1,p}, \delta_{i1}, \ldots, \delta_{i,L-1}\}^\top.$$

Averaging over units, we can write the vector of estimands as

$$\boldsymbol{\theta} = \sum_{\ell=1}^{L} \boldsymbol{Q}_\ell \overline{\boldsymbol{U}}(\boldsymbol{z}_\ell),$$

$$\boldsymbol{\theta} = \{\overline{\tau}_1, \ldots, \overline{\tau}_{L-1}, \overline{\tau}_{1,p}, \ldots, \overline{\tau}_{L-1,p}, \overline{\delta}_1, \ldots, \overline{\delta}_{L-1}\}^\top.$$

Furthermore, we can write the vector of estimators for these quantities defined above as $\widehat{\boldsymbol{\theta}} = \sum_{\ell=1}^{L} \boldsymbol{Q}_\ell \widehat{\boldsymbol{U}}(\boldsymbol{z}_\ell)$, where the first entry of $\widehat{\boldsymbol{\theta}}$ is $\widehat{\tau}_1$ and the other values are defined similarly. For our particular quantities of interest, we have

$$\boldsymbol{Q}_\ell = \begin{pmatrix} 2^{-(K-1)} g_{1\ell} \mathbf{1}_L^\top & \mathbf{0}_L^\top \\ \vdots & \vdots \\ 2^{-(K-1)} g_{L-1,\ell} \mathbf{1}_L^\top & \mathbf{0}_L^\top \\ 2^{-(K-1)} g_{L-1,\ell} (\boldsymbol{g}_1 \circ \boldsymbol{g}_{L-1})^\top & \mathbf{0}_L^\top \\ \vdots & \vdots \\ 2^{-(K-1)} g_{L-1,\ell} (\boldsymbol{g}_{L-1} \circ \boldsymbol{g}_{L-1})^\top & \mathbf{0}_L^\top \\ \mathbf{0}_L^\top & 2^{-K} g_{1\ell} \boldsymbol{g}_1^\top \\ \vdots & \vdots \\ \mathbf{0}_L^\top & 2^{-K} g_{L-1,\ell} \boldsymbol{g}_{L-1}^\top \end{pmatrix},$$

where the exact formulations of each block come from the previous definitions of the estimands.

To assess the asymptotic distribution of the these estimators, we now define several variance and covariance terms. In particular, let

$$\boldsymbol{S}_\ell^2 = \frac{1}{N-1} \sum_{i=1}^{N} [\boldsymbol{U}_i(\boldsymbol{z}_\ell) - \overline{\boldsymbol{U}}(\boldsymbol{z}_\ell)][\boldsymbol{U}_i(\boldsymbol{z}_\ell) - \overline{\boldsymbol{U}}(\boldsymbol{z}_\ell)]^\top$$

and

$$\boldsymbol{S}_{\boldsymbol{\theta}}^2 = \frac{1}{N-1} \sum_{i=1}^{N} [\boldsymbol{\theta}_i - \boldsymbol{\theta}][\boldsymbol{\theta}_i - \boldsymbol{\theta}]^\top.$$

The first of these, $\boldsymbol{S}_\ell^2$ is the variance of the potential outcomes under treatment assignment $\boldsymbol{z}_\ell$, and the second, $\boldsymbol{S}_{\boldsymbol{\theta}}^2$, is the covariance matrix of the individual-level treatment effects. Note that while $\boldsymbol{S}_\ell^2$ can be identified under the present experimental design, $\boldsymbol{S}_{\boldsymbol{\theta}}^2$ cannot be identified because it would require observing individual-level treatment effects. In particular, we can use the sample variance within each treatment arm to estimate $\boldsymbol{S}_\ell^2$,

$$\boldsymbol{s}_\ell^2 = \frac{1}{N_\ell - 1} \sum_{i: W_{i\ell}=1} \left\{\boldsymbol{U}_i - \widehat{\boldsymbol{U}}(\boldsymbol{z}_\ell)\right\} \left\{\boldsymbol{U}_i - \widehat{\boldsymbol{U}}(\boldsymbol{z}_\ell)\right\}^\top.$$

Under Assumption 1 and over the randomization distribution, $\widehat{\boldsymbol{\theta}}$ has mean $\boldsymbol{\theta}$ and covariance

$$\operatorname{cov}(\widehat{\boldsymbol{\theta}}) = \sum_{\ell=1}^{L} \frac{1}{N_\ell} \boldsymbol{Q}_\ell \boldsymbol{S}_\ell^2 \boldsymbol{Q}_\ell^\top - \frac{1}{N} \boldsymbol{S}_{\boldsymbol{\theta}}^2,$$

by Theorem 3 of Li and Ding (2017). This result is a finite-population result and requires no assumptions on the data generating process of the outcomes.

A conservative estimator for the covariance of $\widehat{\boldsymbol{\theta}}$ can be $\widehat{\boldsymbol{V}} = \sum_{\ell=1}^{L} N_\ell^{-1} \boldsymbol{Q}_\ell \boldsymbol{s}_\ell^2 \boldsymbol{Q}_\ell^\top$. Given the above result, this will overestimate the covariance of $\widehat{\boldsymbol{\theta}}$ by $N^{-1}\boldsymbol{S}^2_{\boldsymbol{\theta}}$. This latter quantity is generally unestimable because estimating it would require observing the joint distribution of different potential outcomes, $\{\boldsymbol{U}_i(\boldsymbol{z}_1), \dots, \boldsymbol{U}_i(\boldsymbol{z}_L)\}$. Under the additional stringent assumption that all of the individual-level effects are additive, $\boldsymbol{S}^2_{\boldsymbol{\theta}}$ will be equal to 0 because the effects do not vary across units. In the IV context, however, additive treatment effects are awkward because they would rule out heterogeneous treatment effects that the compliance framework is designed to address.

3.2. *Asymptotic Distribution Under a Finite-Population Approach*

In this subsection, we take a finite-population approach to asymptotics that treats $\Pi_N = \{\boldsymbol{U}_1(\boldsymbol{z}_1), \dots, \boldsymbol{U}_N(\boldsymbol{z}_L)\}$ as a set of fixed population quantities and all randomness comes from the distribution of $\boldsymbol{Z}_i$. To perform asymptotics in this setting, we embed Π_N into a hypothetical sequence of finite populations that grow in size and investigate the properties of our estimators along that sequence (see Lehmann and D'Abrera 1975; Lehmann 1999; Li and Ding 2017, for more on this approach). We assume that we are in a setting where as N increases, N_ℓ also increases without bound for all ℓ. In particular, we assume that N_ℓ/N has a positive limiting value for all ℓ throughout.

We start by getting a consistency result.

Theorem 1 (Consistency). Under Assumption 1 and the assumption that $(1 - N_\ell/N)\boldsymbol{S}^2_\ell/N_\ell \to 0$ as $N \to \infty$, $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \xrightarrow{p} 0$ as $N \to \infty$.

Proof. From finite population results in, for instance, Rosén (1964) and Scott and Wu (1981), the assumption that $(1 - N_\ell/N)\boldsymbol{S}^2_\ell/N_\ell \to 0$ gives us that $\widehat{\boldsymbol{U}}(\boldsymbol{z}) - \overline{\boldsymbol{U}}(\boldsymbol{z}) \xrightarrow{p} 0$ as $N \to \infty$ for all $\boldsymbol{z}$. Therefore,

$$\begin{aligned}\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} &= \sum_{\ell=1}^{L} \mathbf{Q}_\ell \widehat{\boldsymbol{U}}(\boldsymbol{z}_\ell) - \sum_{\ell=1}^{L} \mathbf{Q}_\ell \overline{\boldsymbol{U}}(\boldsymbol{z}_\ell) \\ &= \sum_{\ell=1}^{L} \mathbf{Q}_\ell \left(\widehat{\boldsymbol{U}}(\boldsymbol{z}_\ell) - \overline{\boldsymbol{U}}(\boldsymbol{z}_\ell)\right) \xrightarrow{p} 0.\end{aligned}$$

□

We now move on to distributional results. In order to conduct inference on $\widehat{\boldsymbol{\theta}}$, we need to know not only its moments, but also its distribution. While it is possible to computationally approximate the randomization distribution of $\widehat{\boldsymbol{\theta}}$ under a null hypothesis about $\boldsymbol{\theta}$, this approach can be quite complicated and even infeasible when entertaining non-sharp null hypotheses (Kang, Peck, and Keele 2018). Instead, we rely on finite-population asymptotics to derive an approximation of the distribution $\widehat{\boldsymbol{\theta}}$ as in Li and Ding (2017) and Kang, Peck, and Keele (2018). In this framework, we can derive asymptotic normality of our estimators under a limitation on how much a unit can dominate the population variance. In particular, define the maximum squared distance of the qth coordinate of $\mathbf{Q}_\ell \boldsymbol{U}_i(\boldsymbol{z}_\ell)$ from its population mean,

$$m_\ell(q) = \max_{1 \le i \le N} \left[\mathbf{Q}_\ell \boldsymbol{U}_i(\boldsymbol{z}_\ell) - \mathbf{Q}_\ell \overline{\boldsymbol{U}}(\boldsymbol{z}_\ell)\right]_q^2 \quad 1 \le q \le r,$$

the finite-population variance of the qth coordinate of $\mathbf{Q}_\ell \boldsymbol{U}_i(\boldsymbol{z}_\ell)$,

$$v_\ell(q) = \frac{1}{N-1} \sum_{i=1}^{N} \left[\mathbf{Q}_\ell \boldsymbol{U}_i(\boldsymbol{z}_\ell) - \mathbf{Q}_\ell \overline{\boldsymbol{U}}(\boldsymbol{z})_\ell\right]_q^2 \quad 1 \le q \le r,$$

and the finite-population variance of the qth coordinate of $\boldsymbol{\theta}$,

$$v_{\boldsymbol{\theta}}(q) = \frac{1}{N-1} \sum_{i=1}^{N} [\boldsymbol{\theta}_i - \boldsymbol{\theta}]_q^2 \quad 1 \le q \le r.$$

Li and Ding (2017) derive the following assumptions that are sufficient for asymptotic normality.

Assumption 5. As $N \to \infty$,

$$\max_\ell \max_{1 \le q \le r} \frac{1}{N_\ell^2} \frac{m_\ell(q)}{\sum_{\ell'} N_{\ell'}^{-1} v_{\ell'}(q) - N^{-1} v_{\boldsymbol{\theta}}(q)} \to 0$$

Roughly speaking, this assumption limits how a particular unit can dominate the variance of $\mathbf{Q}_\ell \boldsymbol{U}_i(\boldsymbol{z}_\ell)$, uniformly across all assignment vectors and components of $\boldsymbol{\theta}$. While this assumption is general and difficult to interpret, Li and Ding (2017) demonstrate several more interpretable conditions that imply this assumption. Finally, we impose a regularity condition on the correlation matrix of $\widehat{\boldsymbol{\theta}}$ and derive the asymptotic distribution of the (standardized) ITT estimators.

Assumption 6. The correlation matrix of $\widehat{\boldsymbol{\theta}}$ has limiting value $\boldsymbol{\Sigma}$.

Lemma 1. Under Assumption 1, 5 and 6, by Theorem 4 of Li and Ding (2017), we have

$$\left(\frac{\widehat{\tau}_1 - \overline{\tau}_1}{\sqrt{\operatorname{var}(\widehat{\tau}_1)}}, \dots, \frac{\widehat{\tau}_{L-1} - \overline{\tau}_{L-1}}{\sqrt{\operatorname{var}(\widehat{\tau}_{L-1})}}, \frac{\widehat{\tau}_{1,p} - \overline{\tau}_{1,p}}{\sqrt{\operatorname{var}(\widehat{\tau}_{1,p})}}, \dots, \frac{\widehat{\tau}_{L-1,p} - \overline{\tau}_{L-1,p}}{\sqrt{\operatorname{var}(\widehat{\tau}_{L-1,p})}}, \frac{\widehat{\delta}_1 - \overline{\delta}_1}{\sqrt{\operatorname{var}(\widehat{\delta}_1)}}, \dots, \frac{\widehat{\delta}_{L-1} - \overline{\delta}_{L-1}}{\sqrt{\operatorname{var}(\widehat{\delta}_{L-1})}}\right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}).$$

These results do not rely on any of the instrumental variable assumptions (monotonicity and the exclusion restrictions), and so we can conduct inference on these quantities as ITT effects even if the IV assumptions are suspect. These quantities will gain the additional interpretations in terms of complier effects, as discussed earlier, if the IV assumptions hold.

To get an asymptotic, finite-population distributional result for our IV estimators, which are all ratio estimators, we can use a finite-population delta method (Pashley 2019).

Lemma 2. Under Assumption 1, 5, and 6, assumptions for Theorem 1, and also assuming that $\overline{\delta}_j$ has a nonzero limiting value, we have the following asymptotic normality result for our MCAFE estimators:

$$\frac{\widehat{\phi}_j - \overline{\phi}_j}{\sqrt{\frac{1}{\overline{\delta}_j^2}\mathrm{var}(\widehat{\tau}_j) + \overline{\phi}_j^2\frac{1}{\overline{\delta}_j^2}\mathrm{var}(\widehat{\delta}_j) - 2\overline{\phi}_j\frac{1}{\overline{\delta}_j^2}\mathrm{cov}(\widehat{\tau}_j, \widehat{\delta}_j)}} \xrightarrow{d} \mathrm{N}(0, 1).$$

It is straightforward to extend this result to the PCAFEs. Although the delta method is typically associated with a superpopulation perspective, this is a finite-population asymptotic result only requiring standard assumptions on the asymptotic variance and that $\overline{\delta}_j$ has a nonzero limiting value, which under monotonicity is the same as assuming that the proportion of compliers for that particular effect has a nonzero limiting value. We can construct confidence intervals directly from this distribution by estimating the variance as $\frac{1}{\widehat{\delta}_j^2}\widehat{\mathrm{var}}(\widehat{\tau}_j) + \widehat{\phi}_j^2\frac{1}{\widehat{\delta}_j^2}\widehat{\mathrm{var}}(\widehat{\delta}_j) - 2\widehat{\phi}_j\frac{1}{\widehat{\delta}_j^2}\widehat{\mathrm{cov}}(\widehat{\tau}_j, \widehat{\delta}_j)$. However, we employ a useful trick in the next section to create intervals with potential benefits in terms of coverage and behavior with small compliance probabilities.

Before moving on to this method, we give a final consistency result for our ratio estimators:

Lemma 3. Assume either of the following two sets of conditions:

(a) the assumptions of Theorem 1 and the additional assumptions that all components of $\boldsymbol{\theta}$ have finite limiting values, and in particular nonzero limiting values for the $\overline{\delta}_j$; OR
(b) the assumptions of Lemma 2 and the assumption that $\boldsymbol{S}_\ell^2$ and $\boldsymbol{S}_{\boldsymbol{\theta}}^2$ have finite limiting values.

Then $\widehat{\phi}_j - \overline{\phi}_j \xrightarrow{p} 0$ and $\widehat{\gamma}_j - \overline{\gamma}_j \xrightarrow{p} 0$ as $N \to \infty$, for all $j \in \{1, \ldots, L-1\}$.

Lemma 3 requires additional regularity conditions on the sequence of finite populations beyond those required in Theorem 1 to avoid situations where the ratio of the population ITTs diverges as $N \to \infty$. We provide a proof of this result in supplemental material A.

3.3. Constructing Confidence Intervals for IV Effects: Fieller's Method

The results of the previous section can be used directly to generate confidence intervals. Here we present a method to create intervals originally from Fieller (1954) and used in Kang, Peck, and Keele (2018) and Li and Ding (2017) in the context of instrumental variables, which performs better with low rates of compliance. We can begin from the result of Lemma 2 to derive this method but it is traditional instead to consider the hypothesis test of a particular value, $H_0 : \overline{\phi}_j = \phi_{j0}$, which can be rewritten as $H_0 : \overline{\tau}_j - \phi_{j0}\overline{\delta}_j = 0$. Following Fieller (1954) and Kang, Peck, and Keele (2018), we use the following test statistic to assess this hypothesis:,

$$T(\phi_{j0}) = \widehat{\tau}_j - \phi_{j0}\widehat{\delta}_j.$$

We can use the above asymptotic results to derive the (asymptotic) variance of this statistic as

$$\sigma^2(\phi_{j0}) = \mathrm{var}(\widehat{\tau}_j) + \phi_{j0}^2\mathrm{var}(\widehat{\delta}_j) - 2\phi_{j0}\mathrm{cov}(\widehat{\tau}_j, \widehat{\delta}_j).$$

We can then obtain $\widehat{\mathrm{var}}(\widehat{\tau}_j)$, $\widehat{\mathrm{var}}(\widehat{\delta}_j)$, and $\widehat{\mathrm{cov}}(\widehat{\tau}_j, \widehat{\delta}_j)$ from $\widehat{\boldsymbol{V}}$ for all j and create the following estimator for the variance of the test statistic:

$$\widehat{\sigma}^2(\phi_{j0}) = \widehat{\mathrm{var}}(\widehat{\tau}_j) + \phi_{j0}^2\widehat{\mathrm{var}}(\widehat{\delta}_j) - 2\phi_{j0}\widehat{\mathrm{cov}}(\widehat{\tau}_j, \widehat{\delta}_j).$$

Under the above results about the approximate normality of these quantities, the typical way to assess this hypothesis is to reject the null if $|T(\phi_{j0})/\widehat{\sigma}(\phi_{j0})| \geq z_{1-\alpha/2}$ for some prespecified choice of α. We could then construct a $1-\alpha$ confidence interval for this quantity by inverting the test:

$$\left\{\phi_{j0} : \left|\frac{T(\phi_{j0})}{\widehat{\sigma}(\phi_{j0})}\right| \leq z_{1-\alpha/2}\right\} = \left\{\phi_{j0} : T(\phi_{j0})^2 \leq \widehat{\sigma}^2(\phi_{j0})z_{1-\alpha/2}^2\right\}.$$

Noting that $T(\phi_{j0})^2 = (\widehat{\tau}_j^2 - 2\phi_{j0}\widehat{\tau}_j\widehat{\delta}_j + \phi_{j0}^2\widehat{\delta}_j^2)$, this implies that we can generate the $1-\alpha$ confidence interval by finding: $\{\phi_{j0} : a\phi_{j0}^2 + b\phi_{j0} + c < 0\}$, where

$$\begin{aligned} a &= \widehat{\delta}_j^2 - z_{1-\alpha/2}^2\widehat{\mathrm{var}}(\widehat{\delta}_j) \\ b &= -2\left(\widehat{\tau}_j\widehat{\delta}_j - z_{1-\alpha/2}^2\widehat{\mathrm{cov}}(\widehat{\tau}_j, \widehat{\delta}_j)\right) \\ c &= \widehat{\tau}_j^2 - z_{1-\alpha/2}^2\widehat{\mathrm{var}}(\widehat{\tau}_j). \end{aligned}$$

As in the case of Fieller (1954), Li and Ding (2017), and Kang, Peck, and Keele (2018), the type of interval generated by this quadratic inequality can take several forms: closed interval, disjoint union of tail intervals, or an infinite-length interval that covers the real line. A similar derivation holds for hypotheses about the perfect complier effects, $\overline{\gamma}_j$, replacing $\widehat{\tau}_j$ with $\widehat{\tau}_{j,p}$.

3.4. Inference Under a Superpopulation Model

If we assume that the data are random samples from an infinite superpopulation, some aspects of inference become simpler. In particular, we can view the observations of Y_i with $\boldsymbol{Z}_i = \boldsymbol{z}$ to be a random sample from the superpopulation distribution of $Y_i(\boldsymbol{z})$, independent from the samples of the other treatment assignments. Then, under mild regularity conditions $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to $N(\boldsymbol{0}, \boldsymbol{V})$, where $\boldsymbol{V}$ is the superpopulation variance of $\widehat{\boldsymbol{\theta}}$, and $\widehat{\boldsymbol{V}}$ is a consistent estimator for the asymptotic covariance of $\widehat{\boldsymbol{\theta}}$. One can derive confidence intervals for the superpopulation parameters using $\widehat{\boldsymbol{V}}$ and applying either the above delta method or test-inversion methods.

In supplemental material D we describe a Bayesian approach to inference in this setting as that is a popular way to study both factorial experiments (Dasgupta, Pillai, and Rubin 2015) and instrumental variables (Imbens and Rubin 1997).

4. Comparing Compliance Types

One complication of the factorial setting with noncompliance is the multitude of possible compliance types. We discussed earlier how this made comparing MCAFEs difficult because the underlying compliance group changes from one effect to the next. The solution of focusing on perfect compliers typically has the disadvantage of more variable estimates due to restricting the estimates to a smaller compliance group. In this section, we suggest an alternative path for comparing compliance types: through their possibly varying covariate distributions. We provide two ways of making these comparisons. First, we investigate how the distribution of the covariates changes across different compliance groups. Second, we show one method for adjusting each of the MCAFEs for differences in the distribution of the covariates. We show that under very strong assumptions, the latter can be justified as generalizing from complier-specific effects to the entire sample.

4.1. Covariate Profiles of the Compliance Groups

A common approach to analyzing complier average treatment effects is to profile the compliers in terms of background characteristics. In settings with a single treatment factor, Abadie (2003) showed how to identify the expectations of arbitrary functions of covariates among the compliers. We extend those ideas to the factorial setting.

Let X_i be a vector of observed covariates and $\nu(X_i)$ be a known scalar function of those covariates. We now define an alternative ITT on the product of the this function and the factorial treatment uptake variables, $\widetilde{\boldsymbol{D}}_{ij}$:

$$\bar{\delta}_j(\nu(X_i)) = \frac{1}{N}\sum_{i=1}^{N} 2^{-K}\boldsymbol{g}_j^{\top}\,\nu(X_i)\widetilde{\boldsymbol{D}}_{ij}(\bullet).$$

By similar arguments to ITT on treatment uptake, we can show that

$$\bar{\delta}_j(\nu(X_i)) = \left(\frac{1}{N_j^c}\sum_{i=1}^{N} C_{ij}\nu(X_i)\right)\bar{\delta}_j,$$

so this ITT is the mean of $\nu(X_i)$ among the marginal compliers for effect j multiplied by the proportion of those marginal compliers. Thus, we can recover the means of functions of covariates in the compliance groups with $\bar{\delta}_j(\nu(X_i))/\bar{\delta}_j$. To obtain estimates in our observed samples, we simply replace each of these population quantities with their sample counterparts.

In our empirical example, we use this approach to show how the means of various covariates in each compliance group compare to the overall finite population. Of course, it is straightforward to make these comparisons based on higher moments with the correct choice of $\nu(\cdot)$.

4.2. Adjusting Complier Effects with Compliance Weights

The previous method will allow us to compare the covariate profiles of each compliance group, but this does not give us direct information on how these differences translate into different effects. We now describe one method for putting all the MCAFEs on a similar footing by reweighting them to have the same covariate distribution. We hope that after this reweighting, any remaining variation in estimated effects is not due to compositional differences in compliance groups on the observed covariates. Under a much stronger (and often implausible) generalizability assumption, this procedure will estimate the average factorials effects if compliance (for the given active factors) was forced for all units. These ideas build on the inverse compliance score weighting approach of Aronow and Carnegie (2013), who used a similar methodology to generalize the local average treatment effect (LATE) to the average treatment effect (ATE) in settings with $K = 1$.

We describe the method using the superpopulation framework, to make the notation and interpretation simpler. We define the following compliance weights:

$$\omega_j(x) = \frac{\mathbb{P}(C_{ij} = c)}{\mathbb{P}(C_{ij} = c|X_i = x)},$$

which are inversely proportional to the probability of being a marginal complier for effect j conditional on $X_i = x$. Let $\omega_{ij} = \omega_j(X_i)$. Then, it is straightforward to show that

$$\mathbb{E}[\omega_{ij}\tau_{ij}|C_{ij} = c] = \sum_{x}\mathbb{E}[\tau_{ij}|C_{ij} = c, X_i = x]\mathbb{P}(X_i = x).$$

This shows that the jth weighted MCAFE is a weighted average of conditional MCAFEs where the weights are based on the population distribution of the covariates, not the marginal compliers distribution of the covariates. Thus, we have adjusted for compositional differences related to the covariate distributions in each underlying MCAFE compliance group.

But without further assumptions the conditional MCAFEs are still not comparable because the compliance groups are different for different effects. We can make an additional assumption that will make the weighted MCAFEs comparable. Let τ_{ij}^* be the jth factorial effect for unit i if they were forced to comply with treatment assignment for the active factors in the jth effect regardless of their natural compliance type. Then, we define the following *latent ignorability of compliance* assumption as

$$\mathbb{E}[\tau_{ij}^*|X_i = x] = \mathbb{E}[\tau_{ij}|C_{ij} = c, X_i = x]. \tag{1}$$

This assumption says that for units with the same values of the covariates, the average factorial effect among marginal compliers is the same as the average factorial effect if everyone with $X_i = x$ were forced to comply with the active factors in the jth effect. This assumption is quite strong and may be implausible in many settings. To gain additional intuition, we can use two assumptions which together are stronger but more interpretable. Let $\boldsymbol{T}_{i,j}^*$ be the vector of length K indicating the compliance type for unit i if they are forced to comply for active factors in effect j. Then, for any $\boldsymbol{t}$ such that $t_k = c$ for all $k \in \mathcal{K}(j)$, two sufficient assumptions for 1 are

$$\mathbb{P}[\boldsymbol{T}_{i,j}^* = \boldsymbol{t}|X_i = x] = \mathbb{P}[\boldsymbol{T}_i = \boldsymbol{t}|C_{ij} = c, X_i = x], \tag{2}$$

$$\mathbb{E}[\tau_{ij}^*|\boldsymbol{T}_{i,j}^* = \boldsymbol{t}, X_i = x] = \mathbb{E}[\tau_{ij}|C_{ij} = c, \boldsymbol{T}_i = \boldsymbol{t}, X_i = x]. \tag{3}$$

Assumption (2) says that if we can force noncompliers for the active factors in effect j to comply on those factors, then the distribution of their full compliance types will be the same as the distribution for those who naturally comply to on those factors, conditional on X_i. Assumption (3) says that units with the same full compliance type with forced (or natural) compliance for active factors in effect j will have the same average factorial effect for factor j as those who would naturally comply, conditional on X_i. These assumptions require considerable stability in compliance types and effects across the natural and forced compliance settings that may be difficult to sustain in many applications.

In practice, we can estimate these weights by replacing the population quantities with their sample counterparts. In the empirical application, we stratify the units based on a discrete set of covariates and estimate all quantities within these strata. Aronow and Carnegie (2013) present a parametric approach to estimating these weights when $K = 1$, which could be extended to our setting as well.

5. Empirical Application: The Effect of Political Canvassing on Voter Turnout

A large literature in political science uses field experiments to examine the effectiveness of various strategies for encouraging voter turnout in elections. These strategies include phone calls, door-to-door canvassing, mailers, and more. A ubiquitous problem with these field experiments is noncompliance because relatively few people are willing and able to speak with political canvassers on the phone or at the doorstep. We apply the above framework to a particular get-out-the-vote field experiment fielded in New Haven ahead of the 1998 general election in New Haven, CT (Gerber and Green 2000). In the original experiment, $N = 23,450$ households were randomly assigned three factors: a door-to-door canvassing visit (or not), a phone call (or not), and a mailer sent to their home (or not). Note that door-to-door canvasing was randomized independently of the other two factors, so we are performing a conditional analysis when analyzing as a factorial design, conditioning on the number of people actually assigned to each treatment combination. All of the factors involved messages that encouraged voter turnout. Randomization was done at the household level and the outcome is whether anyone in the household voted in the 1998 general election. Previous studies have analyzed various aspects of this experiment, both substantively and methodologically (Gerber and Green 2000; Imai 2005; Hansen and Bowers 2009; Blackwell 2017).

Noncompliance in this voter mobilization setting usually occurs when a resident fails to answer the door for an in-person canvassing attempt or fails to answer the phone for a phone canvassing attempt. The MCAFE for in-person canvassing, then, would be the effect of canvassing among individuals that would answer their door and talk to a canvasser regardless of whether or not they would answer a phone call or read a mailer. That is, it marginalizes or averages over the assignment for the phone and mailer factors ignoring the actual uptake on those factors. The PCAFE, on the other hand, would be the effect of in-person-canvassing among those who would answer their door, answer their phone, and read any mailer sent to them. In this case, by averaging over the assignment to the other factors, we are also directly averaging over uptake. While noncompliance on the mailers factor is theoretically possible, it is difficult to measure—we would have to know if a person both received the mailer and read it closely enough to get the message. Thus, for the purposes of this application, we assume perfect compliance on the mailers factor. One advantage of our approach is that all estimands, estimators, and confidence intervals are well-defined even when some of the factors have perfect compliance. It also emphasizes the benefits of our MCAFE quantities which can be calculated on any given factor without knowing compliance information for other factors. We estimate that the marginal compliance rates for in-person and phone canvassing is 0.296 and 0.282, respectively. The perfect compliance rate, on the other hand, is estimated as just 0.104.

Figure 1 shows the estimated MCAFEs and PCAFEs for this voter mobilization study with 95% confidence intervals using the Fieller method. The main substantive takeaway from the results is that only in-person canvassing appears to have a positive and statistically significant effect on turnout, at least for marginal compliers. Other MCAFEs, while sometimes having large point estimates, all have confidence intervals that include 0. The effects for perfect compliers also all have confidence intervals that include zero, and all of these intervals are much wider than for marginal compliers. This demonstrates the loss of precision when attempting to make inferences about a smaller group, even if the resulting coefficients are more directly comparable. Even with that increase in uncertainty, there are striking differences between the point estimates of the PCAFEs and MCAFEs, which could also reflect how the perfect compliers in this setting might be behavioral outliers. Given that the in-person canvassing was done during the day, these are people who are home and willing to talk about political campaigns in person or over the phone. We may expect these individuals to have different responses to canvasing attempts than the population at large.

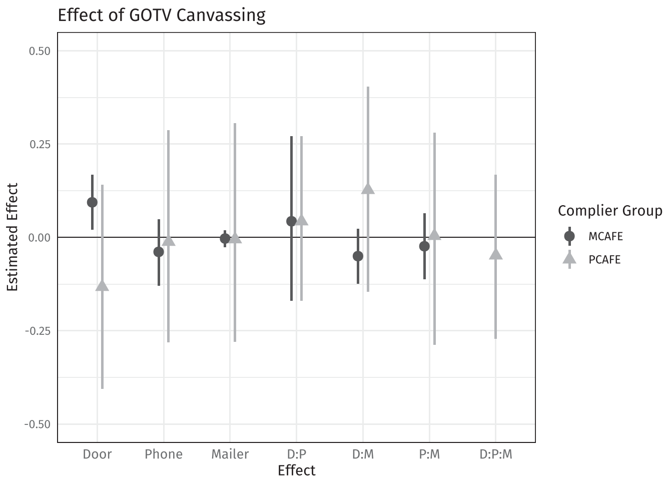

Figure 1. Estimated marginal and perfect complier factorial effects of canvassing methods on voter turnout.

In Figure 2, we use the methods of Section 4 to investigate how these compliance groups and their associated effects relate to background characteristics of the subjects. We have limited data on the households in this study, but we do have average age in the household, household size (in terms of number of registered voters), whether anyone in the house is registered with the Democratic or Republican party, and whether anyone in the household voted in the previous election. The left panel of Figure 2 uses the approach of Section 4.1 and shows the estimated means of these covariates relative to overall sample, and it is clear that compliance with any of the factors is associated with older subjects, more registered voters in the household, and higher rates of previous turnout. These differences appear to be stronger for the phone compliers and the combined door-to-door and phone compliers. This helps explain why the PCAFE and MCAFE point estimates are more disparate for estimating the effect of the door-to-door intervention than for the phone intervention; the subpopulation for which we are estimating the MCAFE for the door-to-door intervention is estimated to be younger, from smaller households, and less likely to have voted previously than the subpopulation for the corresponding PCAFE. And, of course, differences may also exist between these groups on other unmeasured covariates. This exemplifies the heterogeneity in effects we might expect among the

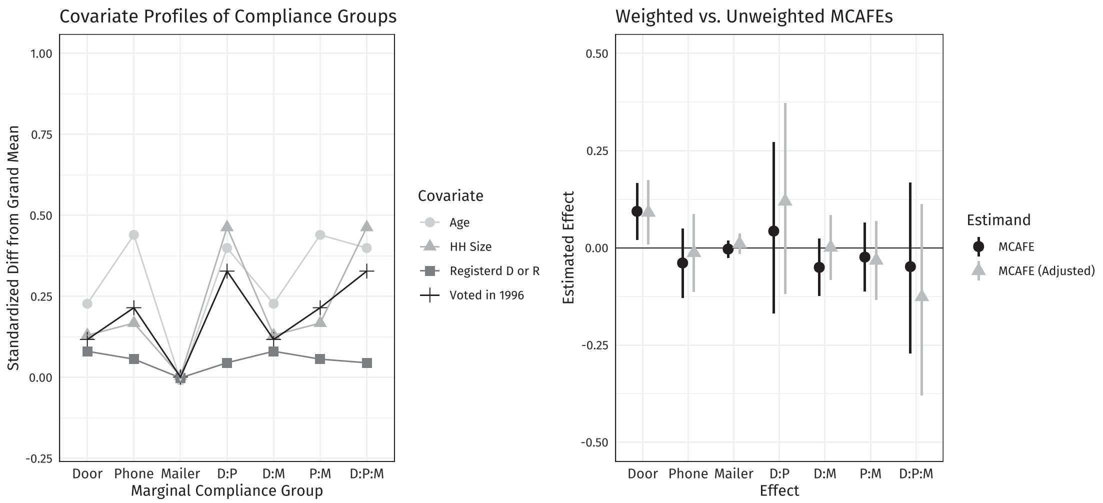

Figure 2. Comparison of estimated covariate means within compliance groups (left) and the estimated MCAFEs using weights to adjust for covariate differences between the compliance groups (right).

different compliance groups. This result also emphasizes that the MCAFEs for different effects are not directly comparable because they relate to different subpopulations, so we are estimating not only the effect of different interventions but we are also averaging over different types of individuals. For instance, the subpopulation corresponding to the MCAFE for the phone intervention is estimated to be almost 0.25 standard deviations older on average than the subpopulation corresponding to the MCAFE for the door-to-door intervention.

As a way to potentially adjust for these covariate differences, we use the weighting approach of Section 4.2. We create a binned version of age and create strata based on the unique values of all the covariates, allowing us to estimate the weights nonparametrically by stratification. The left panel of Figure 2 shows how the weighted MCAFEs compare to the original MCAFEs, with the confidence intervals of the weighted MCAFEs obtained by conditioning on the weights. Small differences between the weighted and unweighted MCAFEs do appear, but the overall substantive conclusions remain unchanged. This provides some evidence that these covariates are not enough to explain the differences we observe, for instance, between the MCAFEs and PCAFEs. We would urge caution in interpreting these weighted MCAFEs as the generalizability assumption needed to allow for comparing effects may not be plausible in this setting.

6. Conclusion

In this article we have presented a new framework for 2^K factorial experiments with noncompliance on any number of factors. Under standard instrumental variable assumptions and a treatment exclusion restriction unique to this setting, we showed how there are several ways to define compliance and we exploited this to define two broad classes of factorial effects: those for marginal compliers and those for perfect compliers. Furthermore, we detailed several ways to estimate and make inferences about these quantities of interest.

There are several avenues for extending this framework. The first would be to consider how to proceed with the identification and estimation of bounds for either the overall average factorial effect or various complier factorial effects when the assumptions maintained in this article do not hold. In particular, the treatment exclusion restriction assumption can be restrictive in that it rules out many types of interactions for compliance. This is especially limiting because interactions are often the target of inference in factorial experiments. Another way to extend this setting would be to allow for more than two levels for each factor given these types of designs are quite common in the social and biomedical sciences. Finally, there are many situations where the compliance status is unknown or only known for a subset of individuals, as in the mailers in the GOTV New Haven experiment. In these settings, it would be useful to use partial identification and bounds to understand what can be learned about the effect of treatment uptake.

Supplementary Materials

The supplementary material contain the following: (A) Proofs and technical notes (B) Alternative estimands and estimators with different weighting (C) Discussion of a weaker treatment exclusion restriction (D) Discussion of how to conduct Bayesian inference (E) Simulation results (F) An additional empirical example.

Acknowledgments

Thanks to Thad Dunning, Kosuke Imai, Luke Keele, Xinran Li, Joel Middleton, and Teppei Yamamoto for comments. Any errors remain our own. Software to implement the methods of this article are available as the `factiv` package for `R` at *https://cran.r-project.org/package=factiv*.

Funding

Nicole Pashley was funded by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303 while working on this article. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

ORCID

Nicole E. Pashley http://orcid.org/0000-0002-5900-3535

References

Abadie, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263. [10]

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. [1,3]

Aronow, P. M., and Carnegie, A. (2013), "Beyond LATE: Estimation of the Average Treatment Effect With an Instrumental Variable," *Political Analysis* 21, 492–506. [10]

Blackwell, M. (2017), "Instrumental Variable Methods for Conditional Effects and Causal Interaction in Voter Mobilization Experiments," *Journal of the American Statistical Association*, 112, 590–599. [1,3,11]

Blattman, C., Jamison, J. C., and Sheridan, M. (2017), "Reducing Crime and Violence: Experimental Evidence From Cognitive Behavioral Therapy in Liberia," *American Economic Review*, 107, 1165–1206. [2]

Cheng, J., and Small, D. S. (2006), "Bounds on Causal Effects in Three-Arm Trials With Non-Compliance," *Journal of the Royal Statistical Society*, Series B, 68, 815–836. [1]

Dasgupta, Tirthankar, Natesh S. Pillai and Donald B. Rubin. 2015. "Causal Inference From 2^K Factorial Designs by Using Potential Outcomes," *Journal of the Royal Statistical Society*, Series B, 77, 727–753. [1,2,3,9]

de la Cuesta, B., Egami, N., and Imai, K. (2021), "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution." *Political Analysis*, pp. 1–27. Available online. [4]

Egami, N., and Imai, K. (2019), "Causal Interaction in Factorial Experiments: Application to Conjoint Analysis," *Journal of the American Statistical Association* 114, 529–540. [4]

Fieller, E. C. (1954), "Some Problems in Interval Estimation," *Journal of the Royal Statistical Society* , Series B, 16, 175–185. [2,7,9]

Fisher, R. A. (1935), *The Design of Experiments,* Edinburgh: Oliver and Boyd. [1,7]

Gerber, A. S., and Green, D. P. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment," *American Political Science Review*, 94, 653–663. [2,11]

Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2014), "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices Via Stated Preference Experiments," *Political Analysis*, 22, 1–30. [1,4]

Hansen, B. B., and Bowers, J. (2009), "Attributing Effects to a Cluster-Randomized Get-Out-the-Vote Campaign," *Journal of the American Statistical Association*, 104, 873–885. [11]

Imai, K. (2005), "Do Get-Out-the-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments," *American Political Science Review*, 99, 283–300. [11]

Imbens, G. W., and Rubin, D. B. (1997), "Bayesian Inference for Causal Effects in Randomized Experiments With Noncompliance," *The Annals of Statistics*, 25, 305–327. [9]

Imbens, G. W., and Rosenbaum, P. R. (2005), "Robust, Accurate Confidence Intervals With a Weak Instrument: Quarter of Birth and Education," *Journal of the Royal Statistical Society*, Series A, 168, 109–126. [7]

Kang, H., Peck, L., and Keele, L. (2018), "Inference for Instrumental Variables: A Randomization Inference Approach," *Journal of the Royal Statistical Society*, Series A, 181, 1231–1254. [2,8,9]

Lehmann, E. L. (1999), *Elements of Large Sample Theory*, New York: Springer. [8]

Lehmann, E. L., and D'Abrera, H. J. M. (1975). *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco, CA: Holden-Day, Inc. [8]

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. [2,7,8,9]

Montgomery, D. C. (2013), *Design And Analysis of Experiments*, 8th ed. New York: Wiley. [1]

Pashley, N. E. (2019), "Note on the Delta Method for Finite Population Inference with Applications to Causal Inference." Available at: https://arxiv.org/abs/1910.09062. [8]

Robins, J. M. (1989), "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, eds. A. Mulley, L. Sechrest, and H. Freeman. Washington, DC: U.S. Public Health Service, National Center for Health Services Research, pp. 113–159. [3]

Rosén, B. (1964), "Limit Theorems for Sampling From Finite Populations," *Arkiv för Matematik* 5, 383–424. [8]

Rubin, D. B. (1980), "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment," *Journal of the American Statistical Association* 75, 591–593. [2]

Schochet, P. Z. (2020), "The Complier Average Causal Effect Parameter for Multiarmed RCTs," *Evaluation Review*, 44, 410–436. [1]

Scott, A., and Wu, C.-F. (1981), "On the Asymptotic Distribution of Ratio and Regression Estimators," *Journal of the American Statistical Association* 76, 98–102. [8]

Wald, A. (1940). "The Fitting of Straight Lines if Both Variables are Subject to Error," *The Annals of Mathematical Statistics*, 11, 284–300. [4]

Yates, F. 1937. "Design and Analysis of Factorial Experiments," *Technical Communication*, 35, Imperial Bureau of Soil Science. [1]