Biometrika (2019), **106**, 2, *pp.* 487–494
Printed in Great Britain

doi: 10.1093/biomet/asy072
Advance Access publication 4 February 2019

Randomization tests of causal effects under interference

By G. W. BASSE

Department of Statistics, University of California, 418 Evans Hall, Berkeley,
California 94720, U.S.A.
gbasse@berkeley.edu

A. FELLER

Goldman School of Public Policy, University of California, 2607 Hearst Avenue,
Berkeley, California 94720, U.S.A.
afeller@berkeley.edu

and P. TOULIS

Booth School of Business, University of Chicago, 5807 South Woodlawn Avenue, Chicago,
Illinois 60637, U.S.A.
panos.toulis@chicagobooth.edu

Summary

Many causal questions involve interactions between units, also known as interference, for example between individuals in households, students in schools, or firms in markets. In this paper we formalize the concept of a conditioning mechanism, which provides a framework for constructing valid and powerful randomization tests under general forms of interference. We describe our framework in the context of two-stage randomized designs and apply our approach to a randomized evaluation of an intervention targeting student absenteeism in the school district of Philadelphia. We show improvements over existing methods in terms of computational and statistical power.

Some key words: Causal inference; Conditional randomization test; Exact test; Interference.

1. Introduction

Classical approaches to causal inference assume that units do not interact with each other, known as the no-interference assumption (Cox, 1958). Many causal questions, however, are inherently about interference between units (Sobel, 2006; Hudgens & Halloran, 2008), and standard approaches often break down. For example, randomization tests on sharp null hypotheses of no effect (Fisher, 1935) are more challenging in the presence of interference because these hypotheses are usually not sharp when there are interactions between units.

Aronow (2012) and Athey et al. (2018) addressed this issue by proposing conditional randomization tests restricted to a subset of units, often called focal units, and a subset of assignments for which the specified null hypothesis is sharp for every focal unit. While the randomization-based approaches in these papers have advantages over model-based approaches (Bowers et al., 2013; Toulis & Kao, 2013), they either explicitly forbid any conditioning that depends on the observed treatment assignment (Athey et al., 2018) or only give limited guidance on how to carry out such conditioning (Aronow, 2012). This constraint may affect testing power because, under interference, realized interactions between units depend on the treatment assignment. The constraint also makes implementing the procedure as a permutation test more difficult, which is an often-neglected practical problem.

In this paper we develop a framework for constructing valid and powerful randomization tests under interference. To do so, we formalize the concept of a conditioning mechanism. The proposed framework enables flexible conditional randomization tests that can condition on the observed treatment assignment. Current methods for randomization tests in the presence of interference are special cases of our framework and correspond to mechanisms that generally fail to leverage the problem structure effectively. For example, current methods often include units whose outcomes provide no information for the null hypothesis of interest, leading to unnecessary loss of power. In our framework, it is straightforward to exclude such units from the test via additional conditioning. Furthermore, more flexible conditioning can typically yield permutation tests that are straightforward to implement, resulting in computational gains.

We apply this approach to two-stage randomized designs, which are often used for assessing causal effects related to interference (Hudgens & Halloran, 2008). First, we show how to apply our framework in this setting by suggesting concrete conditioning mechanisms for various hypotheses. Second, we analyse data from a two-stage randomized evaluation of an intervention targeting student absenteeism in the school district of Philadelphia. Our test is more powerful than alternative methods when applied to the absenteeism study, with a roughly one-third increase in statistical power. Furthermore, our method yields a permutation test on the exposures of interest; alternative methods cannot be implemented as permutation tests.

2. General results for randomization testing

2.1. *Classical randomization tests*

Consider N units indexed by $i = 1, \ldots, N$, and a binary treatment assignment vector $Z \in \{0,1\}^N$, where the ith component, Z_i, is the treatment assignment of unit i. The assignment vector is sampled with probability $\text{pr}(Z)$. Denote by $Y_i(Z)$ the scalar potential outcome of unit i under assignment vector Z. Under the stable unit treatment value assumption (Rubin, 1980), the potential outcome of unit i depends only on its own assignment. Each unit therefore has two potential outcomes, typically denoted as $Y_i(1)$ and $Y_i(0)$, which correspond to outcomes when unit i receives treatment or control, respectively. A classic goal is to test the sharp null hypothesis of zero treatment effect for all units,

$$H_0 : Y_i(1) = Y_i(0) \quad (i = 1, \ldots, N). \tag{1}$$

We can assess H_0 by randomization (Fisher, 1935). Let $T(Z \mid Y)$ denote the test statistic; for example, $T(Z \mid Y) = \text{Ave}(Y_i \mid Z_i = 1) - \text{Ave}(Y_i \mid Z_i = 0)$ is the usual difference in means between treated and control units, where Ave denotes sample average. Let $T^{\text{obs}} = T(Z^{\text{obs}} \mid Y^{\text{obs}})$ denote the observed value of the test statistic, where $Z^{\text{obs}} \sim \text{pr}(Z^{\text{obs}})$ is the observed assignment vector in the experiment, and $Y^{\text{obs}} = Y(Z^{\text{obs}})$ is the corresponding observed outcome vector. Finally, calculate the p-value

$$\text{pval}(Z^{\text{obs}}) = E_Z[\mathbb{I}\{T(Z \mid Y^{\text{obs}}) \geqslant T^{\text{obs}}\}], \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and E_Z is the expectation with respect to the distribution of Z. This test is valid at any level α; that is, $\text{pr}\{\text{pval}(Z^{\text{obs}}) \leqslant \alpha\} \leqslant \alpha$, for all $\alpha \in [0,1]$ when the null hypothesis is true. The key property that ensures validity of (2) is that, under H_0, the value of $T(Z \mid Y)$ can be imputed for every possible counterfactual assignment vector Z', using only outcomes Y^{obs} observed under Z^{obs}. This property allows us to construct the correct sampling distribution of the test statistic. We state the property formally in the following definition, as it will be useful for extending the classical randomization test to settings with interference.

Definition 1. *A test statistic $T(Z \mid Y)$ is imputable with respect to a null hypothesis H_0 if for all Z, Z', for which* $\text{pr}(Z) > 0$ *and* $\text{pr}(Z') > 0$,

$$T\{Z' \mid Y(Z')\} = T\{Z' \mid Y(Z)\}. \tag{3}$$

The key property of an imputable test statistic is that we can simulate its sampling distribution under the null hypothesis H_0, even though we only observe one vector of outcomes, namely $Y(Z^{\text{obs}})$. In the classical setting with no interference, (3) follows from the stable unit treatment value assumption and the sharp null hypothesis in (1), which together imply that $Y(Z') = Y(Z)$, for any possible Z, Z'. Thus, in the classical setting all potential outcomes are imputable, and, by extension, any test statistic is imputable.

2.2. *Randomization tests via conditioning mechanisms*

We now demonstrate that we can obtain valid tests without requiring the stable unit treatment value assumption or a sharp null hypothesis. To do so, we introduce the concept of a conditioning event, $\mathcal{C}$, which is a random variable that is realized in the experiment; we leave this concept abstract for now and give concrete examples below. The key idea is to choose an event space and some conditional distribution $m(\mathcal{C} \mid Z)$ on that space, such that, conditional on $\mathcal{C}$, a test statistic $T(Z \mid Y, \mathcal{C})$ is imputable with respect to the null hypothesis. We refer to $m(\mathcal{C} \mid Z)$ as the conditioning mechanism; $m(\mathcal{C} \mid Z)$ and the design $\text{pr}(Z)$ together induce a joint distribution, $\text{pr}(Z, \mathcal{C}; m) = m(\mathcal{C} \mid Z)\text{pr}(Z)$. With these concepts, we can now state our first main result.

Theorem 1. *Let H_0 be a null hypothesis and $T(Z \mid Y, \mathcal{C})$ a test statistic, such that T is imputable with respect to H_0 under some conditioning mechanism $m(\mathcal{C} \mid Z)$; that is, under H_0,*

$$T\{Z' \mid Y(Z'), \mathcal{C}\} = T\{Z' \mid Y(Z), \mathcal{C}\}, \tag{4}$$

for all $Z, Z', \mathcal{C}$ for which $\text{pr}(Z, \mathcal{C}; m) > 0$ and $\text{pr}(Z', \mathcal{C}; m) > 0$. Consider the procedure where we first draw $\mathcal{C} \sim m(\mathcal{C} \mid Z^{\text{obs}})$, and then compute the conditional p-value,

$$\text{pval}(Z^{\text{obs}}; \mathcal{C}) = E_Z[\mathbb{I}\{T(Z \mid Y^{\text{obs}}, \mathcal{C}) > T^{\text{obs}}\} \mid \mathcal{C}], \tag{5}$$

where $T^{\text{obs}} = T(Z^{\text{obs}} \mid Y^{\text{obs}}, \mathcal{C})$, and the expectation is with respect to $\text{pr}(Z \mid \mathcal{C}) = \text{pr}(Z, \mathcal{C}; m)/\text{pr}(\mathcal{C})$. This procedure is valid at any level, that is, $\text{pr}\{\text{pval}(Z^{\text{obs}}; \mathcal{C}) \leqslant \alpha \mid \mathcal{C}\} \leqslant \alpha$, for any $\alpha \in [0, 1]$, under H_0.

Equation (4) is the critical property that the test statistic is imputable, and directly generalizes (3). As before, the key implication of (4) is that we can simulate from the null distribution of $T\{Z \mid Y(Z), \mathcal{C}\}$, given any possible conditioning event $\mathcal{C}$. Theorem 1 allows us to extend conditional randomization testing to more complicated settings, including testing under interference. Before turning to these settings, we briefly demonstrate that classical examples of randomization testing are special cases of Theorem 1.

Example 1. Let the conditioning event space be such that $T(Z \mid Y, \mathcal{C}) \equiv T(Z \mid Y)$ and $\mathcal{C} \perp\!\!\!\perp Z$. Then the procedure in Theorem 1 reduces to the classical Fisher randomization test described in § 2.1.

Example 2. Hennessy et al. (2016) propose a conditional test that adjusts for covariate imbalance, quantified via a function $B(Z, X)$, where X denotes a covariate vector. For instance, B may be the vector of covariate means in each treatment arm, $B(Z, X) = \{\text{Ave}(X_i|Z_i = 1), \text{Ave}(X_i|Z_i = 0)\}$. Let $\text{pr}(Z) = \text{Unif}\{(0, 1)^N\}$ be a Bernoulli randomization design, and consider the conditioning mechanism defined as $\text{pr}(Z, \mathcal{C}) = \mathbb{I}\{B(Z, X) = \mathcal{C}\}\text{pr}(Z)$. Let $T(Z \mid Y, \mathcal{C}) \equiv T(Z \mid Y)$ be independent of $\mathcal{C}$, and let H_0 be as in (1). Then, the procedure of Theorem 1 corresponds exactly to that of Hennessy et al. (2016).

3. Randomization tests for general exposure contrasts

3.1. *General exposure contrasts*

We now turn to constructing valid randomization tests in the presence of interference. Following Manski (2013) and Aronow & Samii (2017), we consider an exposure mapping $h_i(Z) : \{0, 1\}^N \rightarrow \mathcal{H}$, where $\mathcal{H}$ is an arbitrary set of possible treatment exposures equipped with an equality relationship. Given an exposure mapping, a natural assumption that generalizes the classical stable unit treatment value assumption is

$$Y_i(Z) = Y_i(Z') \quad (i = 1, \ldots, N) \text{ for all } Z, Z' \text{ for which } h_i(Z) = h_i(Z'). \tag{6}$$

This assumption states that potential outcomes are functions only of the exposure, rather than of the entire assignment vector. In the most restrictive case of no interference, the exposure mapping is $h_i(Z) = Z_i$; in the most general case without any restrictions on interference, the exposure mapping is $h_i(Z) = Z$. An example of an intermediate case is if $h_i(Z) = \sum_{j \in \mathcal{N}_i} Z_j$, where $\mathcal{N}_i$ is the set of unit i's neighbours in some network between units, and the exposure mapping of i is therefore the number of is treated neighbours (Toulis & Kao, 2013). In these examples, we implicitly defined $\mathcal{H} = \{0, 1\}$, $\mathcal{H} = \{0, 1\}^N$, and $\mathcal{H} = \mathbb{N}$, respectively.

We can now formulate hypothesis tests on contrasts between treatment exposures. Let $\{a, b\} \subseteq \mathcal{H}$ be two exposures of interest. The null hypothesis on the contrast between exposures a and b is

$$H_0: \; Y_i(Z) = Y_i(Z') \quad (i = 1, \ldots, N) \text{ for all } Z, Z' \text{ for which } h_i(Z), h_i(Z') \in \{a, b\}. \tag{7}$$

The classical sharp null hypothesis in (1) is a special case of (7), with $\mathcal{H} = \{a, b\} = \{0, 1\}$. Under the no interference setting of (1), we can permute the vector of unit exposures $\{a, b\}$ by permuting the treatment assignment vector because the null hypothesis contains all possible exposures. In most interference settings, however, the null hypothesis in (7) is not sharp because it only considers a subset of possible exposures. As a result, observing $Y(Z^{\text{obs}})$ gives only limited information about counterfactual outcomes $Y(Z')$, with $Z' \neq Z^{\text{obs}}$. Since h_i may have arbitrary form, we cannot permute unit exposures by naively permuting the treatment assignment vector.

3.2. *Constructing valid tests for general exposure contrasts*

Testing a contrast hypothesis as in (7) is challenging because only a subset of units is exposed to exposures a or b, and only for a subset of assignment vectors. We therefore construct conditioning events in terms of both units and treatment assignment vectors. Specifically, let $\mathbb{C} = \{(\mathcal{U}, \mathcal{Z}) : \mathcal{U} \subseteq \mathbb{U}, \mathcal{Z} \subseteq \mathbb{Z}\}$ be the space of conditioning events, where $\mathbb{U}$ denotes the power set of units, and $\mathbb{Z}$ denotes the power set of assignment vectors. For some conditioning event $\mathcal{C} = (\mathcal{U}, \mathcal{Z}) \in \mathbb{C}$, the conditioning mechanism can be decomposed, without loss of generality, as

$$m(\mathcal{C} \mid Z) = f(\mathcal{U} \mid Z) g(\mathcal{Z} \mid \mathcal{U}, Z), \tag{8}$$

where f and g are distributions over $\mathbb{U}$ and $\mathbb{Z}$, respectively. Given conditioning event $\mathcal{C} = (\mathcal{U}, \mathcal{Z})$, we consider test statistics, $T(Z \mid Y, \mathcal{C})$, that depend only on outcomes of units in $\mathcal{U}$; following terminology in Athey et al. (2018), we call $\mathcal{U}$ the set of focal units. For example, we can set $T(Z \mid Y, \mathcal{C})$ to be the difference in means between focal units exposed to a and units exposed to b:

$$T(Z \mid Y, \mathcal{C}) = \text{Ave}\{Y_i \mid i \in \mathcal{U}, h_i(Z) = a\} - \text{Ave}\{Y_i \mid i \in \mathcal{U}, h_i(Z) = b\}. \tag{9}$$

Theorem 2. *Let H_0 be a null hypothesis as in (7), let $m(\mathcal{C} \mid Z)$ be a conditioning mechanism as in (8), let $\mathcal{C} = (\mathcal{U}, \mathcal{Z})$, and let T be a test statistic defined only on focal units, as in (9). Then, T is imputable under H_0 if $m(\mathcal{C} \mid Z) > 0$ implies that $Z \in \mathcal{Z}$, and for every $i \in \mathcal{U}$ and $Z' \in \mathcal{Z}$, that*

$$h_i(Z') \in \{a, b\}, \quad h_i(Z) \in \{a, b\}, \tag{10}$$

or

$$h_i(Z') = h_i(Z), \quad h_i(Z) \notin \{a, b\}. \tag{11}$$

If T is imputable the randomization test for H_0 described in Theorem 1 is valid at any level α.

Building on Theorem 2, we can construct a family of valid conditional randomization tests by enumerating the assignment vectors for which conditions (10) and (11) hold. As an example, for any choice of $f(\mathcal{U} \mid Z)$ we could define $g(\mathcal{Z} \mid \mathcal{U}, Z)$ as follows:

$$g(\mathcal{Z} \mid \mathcal{U}, Z) = 1, \text{ only if } \mathcal{Z} = \{Z' \in \mathbb{Z} : \text{Equations (10) and (11) are satisfied for } Z'\}. \tag{12}$$

With this definition, g is degenerate, and so the conditioning mechanism $m(\mathcal{C} \mid Z)$ is indexed solely by the conditional distribution, f, of focal units; we denote these conditioning mechanisms $m_{[f]}$. Our methodology provides many possible conditioning mechanisms that yield valid conditional randomization tests. We can select conditioning mechanisms with desired characteristics, such as high power. For example, we can choose f to maximize the expected number of focal units whose outcomes are informative about H_0. We refer to this set of units as the set of effective focal units, $\mathrm{eff}(\mathcal{U})$, where $\mathrm{eff}(\mathcal{U}) = \{i \in \mathcal{U} : h_i(Z^{\mathrm{obs}}) = a \text{ or } b\}$. Similarly, we could ensure that the number of possible randomizations is also large, and maximize the quantity $|\mathrm{eff}(\mathcal{U})||\mathcal{Z}|$. The choices should be tailored to the specific application.

4. Application: interference in two-stage randomized trials

4.1. *Two-stage randomized trials*

We now turn to the use of conditional randomization tests in two-stage randomized trials, which are used to assess spillovers between units (Hudgens & Halloran, 2008). Specifically, we consider the setting of Basse & Feller (2018), in which N units reside in K households indexed by $k = 1, \ldots, K$. In the first stage of the two-stage randomized trial, K_1 households are assigned to treatment, completely at random. In the second stage, one individual in each treated household is assigned to treatment, completely at random. As before, $Z_i \in \{0, 1\}$ is the assignment of unit i, and $Z = (Z_1, \ldots, Z_N)$ is the entire assignment vector. There is a residence index R_{ij}, such that $R_{ij} = 1$ if unit i resides in household j, and is 0 otherwise. Let $[i] = \sum_j jR_{ij}$ denote the household wherein unit i resides. Finally, let $W = (W_1, \ldots, W_K)$ denote the assignment vector on the household level, so that $W_j = \sum_i Z_i R_{ij}$.

The stable unit treatment assumption is not realistic in this context, so we make two assumptions on the interference structure that will imply a specific exposure mapping. First, we make the partial interference assumption (Sobel, 2006): units can interact within, but not between, households. Second, we make the stratified interference assumption (Hudgens & Halloran, 2008): unit i's potential outcomes only depend on the number of units treated in the household, here 0 or 1, rather than the precise identity of the treated unit. Manski (2013) calls this the anonymous interactions assumption. See Hudgens & Halloran (2008) for additional discussion.

These two assumptions can be expressed by the exposure mapping $h_i(Z) = (Z_i, W_{[i]})$. Since the potential outcome of unit i depends only on $h_i(Z)$ by the assumption in equation (6), for brevity we will use $Y_i(Z_i, W_{[i]})$ to denote the value of $Y_i(Z)$. Thus, unit i's potential outcome can take only three values:

$$Y_i(Z) \in \{Y_i(0, 0), Y_i(0, 1), Y_i(1, 1)\},$$

that is, $Y_i(0, 0)$ if unit i is a control unit in a control household; $Y_i(0, 1)$ if unit i is a control unit in a treated household; and $Y_i(1, 1)$ if unit i is a treated unit in a treated household. The fourth combination, $Y_i(0, 1)$, is not possible because when unit i is treated, household $[i]$ is also treated. Thus, the space of exposures is $\mathcal{H} = \{a, b, c\}$, with $a = (0, 0), b = (0, 1), c = (1, 1)$.

4.2. *A valid test for spillovers in two-stage designs*

We now focus on testing the null hypothesis of no spillover effect:

$$H_0^s : \; Y_i(0, 0) = Y_i(0, 1) \quad (i = 1, \ldots, N). \tag{13}$$

The Supplementary Material contains analysis and results for the null hypothesis of no primary effect, $H_0^p : Y_i(0, 0) = Y_i(1, 1)$, for every unit i. Equation (13) is a special case of the exposure contrast as defined in (7), with $a = (0, 0)$ and $b = (0, 1)$. As in § 3.2, we set the test statistic to be the difference in means between the two exposures. The challenge is to find a conditioning mechanism that guarantees validity while preserving power.

We impose two constraints on our choice of focal units. First, units that are exposed to $c = (1, 1)$ are excluded from being selected as focal units because these units do not contribute to the test statistic. Equivalently, we want to exclude units assigned to $Z_i = 1$ from being focal. This is an example of conditioning using observed assignment Z. Second, we choose a single nontreated unit at random from each household as the focal unit. In the Supplementary Material, we show that choosing one focal unit per household leads to a randomization test that is equivalent to a permutation test on the exposures of interest, $a = (0, 0)$ and $b = (0, 1)$, which simplifies computation.

Proposition 1. *Consider the following testing procedure.*

Step 1. In control households ($W_j = 0$), choose one unit at random. In treated households ($W_j = 1$), choose one unit at random among the nontreated units ($Z_i = 0$).
Step 2. Compute the distribution of the test statistic in equation (9) induced by all permutations of exposures on the chosen units, using $a = (0, 0)$ and $b = (0, 1)$ as the contrasted exposures.
Step 3. Compute the p-value.

Steps 1–3 outline a procedure that is valid for testing the null hypothesis of no spillover effect, H_0^s.

We show in the Supplementary Material that the procedure in Proposition 1 is an application of Theorem 2 with a conditioning mechanism defined by

$$f(\mathcal{U} \mid Z) = \mathrm{Un}\left\{\mathcal{U} \subseteq \mathbb{U} \,:\, Z_i\mathbb{I}(i \in \mathcal{U}) = 0,\quad \sum_{i'} \mathbb{I}(i' \in \mathcal{U})R_{i'j} = 1,\quad \text{for every } i,j\right\}. \tag{14}$$

As discussed earlier, the first constraint in (14) ensures that we only select focal units, $i \in \mathcal{U}$, that are not assigned to treatment; the second constraint restricts the focal set to one unit per household.

4.3. *Comparison with existing methods*

Our approach builds on several existing methods. Aronow (2012), who outlines some ideas that we discuss here, develops a test for the null hypothesis of no spillover effect. Although that paper does not exclude conditioning on Z, it gives limited guidance on how such conditioning would work. Athey et al. (2018) extend the method of Aronow (2012) to a broader class of hypotheses, but explicitly forbid the selection of focal units to depend on the realized assignment Z. In the Supplementary Material, we show that their approaches are equivalent to choosing a set of focal units independent of Z; that is, $f(\mathcal{U} \mid Z) \equiv f(\mathcal{U})$. In fact, in the two-stage design we consider, the methods of Aronow (2012) and Athey et al. (2018) are identical; see the Supplementary Material.

Athey et al. (2018) recognize that choosing focal units completely at random often yields tests with low power. They therefore propose more sophisticated approaches for selecting focal units using additional information. For instance, Athey et al. (2018) advocate selecting focal units via ϵ-nets: first select a focal unit, possibly at random, then choose subsequent focal units beyond a graph distance ϵ from that focal unit. In our applied example, this approach suggests choosing one focal unit at random from each household:

$$f(\mathcal{U} \mid Z) \equiv f(\mathcal{U}) = \mathrm{Un}\left\{\mathcal{U} \subseteq \mathbb{U} \,:\, \sum_{i} \mathbb{I}(i \in \mathcal{U})R_{ij} = 1,\ \text{for every } j\right\}. \tag{15}$$

Our proposed design in (14) has two main advantages over the design in (15). First, we can implement our design via a simple permutation test, as described in Proposition 1. This is not always possible for the design in (15). In fact, the Supplementary Material shows that a conditioning mechanism based on (15) is a permutation test only when households have equal size, which does not hold in our application. In the absence of a permutation test, an analyst working with the conditioning mechanism defined by (15) has to calculate the support of g in (12) fully and exactly, and take uniform draws over that set to sample from the correct randomization distribution. This calculation is exponentially hard, and there are no theoretical guarantees for when the test of Athey et al. (2018) can generally be implemented as a simple permutation test.

Second, unlike in our proposed design, the design in (15) may include treated units as focal units. Since treated units are not part of the effective focal set for testing the null hypothesis of no spillover effect, including them will reduce power. In particular, our design will always have at least as many effective focal units as the design in (15), and at least as many assignment vectors in the randomization test. To quantify this, suppose that all households have n units. We show in the Supplementary Material that for the choice of $f(\mathcal{U} \mid Z) = f(\mathcal{U})$ in (15), the number of effective focal units has distribution $|\text{eff}(\mathcal{U})| \sim K - K_1 + \text{Bi}(K_1, 1/n)$, where K is the number of all households, and K_1 is the number of treated households, so $E\{|\text{eff}(\mathcal{U})|\} = K - K_1(1 - 1/n)$. By contrast, the choice of $f(\mathcal{U} \mid Z)$ in (14) leads to a number of effective focal units that is always equal to K, the number of all households. In the experiment we describe next, there are 3169 households with $n = 2$ units. Restricting to this subset, the design in (15) has an average of 2123 effective focal units, a reduction of one-third from our proposed design.

Finally, Rigdon & Hudgens (2015) propose a method for calculating exact confidence intervals in two-stage randomized designs with binary outcomes. However, it is not applicable to our setting with continuous outcome, nor is the proposed approximation well-suited for tests of a given null hypothesis.

4.4. *Application to a school attendance experiment*

We illustrate our approach using a randomized trial of an intervention designed to increase student attendance in the school district of Philadelphia (Rogers & Feller, 2018). Following the set-up in Basse & Feller (2018), we focus on a subset of this experiment with $N = 8654$ students in $K = 3876$ multi-student households, of which $K_1 = 2568$ were treated. For this subset, the district sent targeted attendance information to the parents about only one randomly chosen student in that household. The outcome of interest is the number of days absent during the remainder of the school year. Following Rosenbaum (2002), we focus on regression-adjusted outcomes, adjusting for a vector of pre-treatment covariates, including demographics and prior year attendance. Additional details on the analysis are included in the Supplementary Material, including results for the primary effect.

To assess spillovers, we sample 100 sets $\mathcal{U}^{(l)}$ $(l = 1, \ldots, 100)$ for both ours and Athey et al. (2018)'s choice of function $f(\mathcal{U} \mid Z)$. For each set, we compute p-values for the null hypothesis of no spillover effect H_0^s in (13) and report whether it rejects with $p < 0.05$. Overall, the test using Athey et al. (2018)'s method rejects the null hypothesis of no effect for 66% of focal sets; the test using our method rejects the null hypothesis of no effect for 92% of focal sets.

We also obtain confidence intervals and Hodges–Lehmann point estimates by inverting a sequence of randomization tests under an additive treatment effect model, $Y_i(1, 0) = Y_i(0, 0) + \tau^s$ $(i = 1, \ldots, N)$. For each focal set, obtaining these quantities is straightforward given $\mathcal{U}$ via standard methods (Rosenbaum, 2002). Aggregating information across focal sets, however, remains an open problem; we discuss this briefly in § 5. For simplicity, we summarize the results by presenting medians across focal sets. For our proposed approach, the median value of the Hodges–Lehmann point estimates is $\hat{\tau}^{(\text{cond})} \approx -1$ day, with associated 95% confidence interval $[-1.70, -0.34]$. For the method of Athey et al. (2018), the median point estimate is $\hat{\tau}^{(\text{rand})} \approx -1.1$ days, with associated 95% confidence interval $[-1.84, -0.28]$. Across focal sets, the average width of the confidence intervals obtained via Athey et al. (2018)'s method is 1.60, compared to 1.42 with our approach, a reduction of 11%.

Results from both approaches are in line with those obtained by Basse & Feller (2018) via unbiased estimators. These confirm the presence of substantial within-household spillover effect that is nearly as large as the primary effect, suggesting that intra-household dynamics play a critical role in reducing student absenteeism and should be an important consideration in designing future interventions.

5. Discussion

Constructing appropriate conditioning mechanisms can be challenging in settings more complex than two-stage designs. Doing so requires understanding the interference structure and finding powerful conditioning mechanisms subject to that structure. Furthermore, conditioning mechanisms produce a distribution of p-values across random choices for the conditioning event. While this does not affect the validity of the test, it raises problems such as interpretation and sensitivity of the test results (Geyer & Meeden, 2005). At the same time, the distribution itself may contain information useful to improve the power of the test. In ongoing research, we are working to use multiple testing methods to address this problem.

Acknowledgement

The authors thank the associate editor and two referees for detailed feedback. We also thank Peng Ding, Dean Eckles, Michael Hudgens, Kosuke Imai, Luke Miratrix, Todd Rogers, Fredrik Sävje, John Ternovski, and Teppei Yamamoto for helpful feedback and discussion. Feller gratefully acknowledges financial support from a Spencer Foundation/National Academy of Education postdoctoral fellowship.

Supplementary material

Supplementary material available at *Biometrika* online includes the proofs of all theorems and claims, additional empirical work from § 4.4, and a counterpart to § 4.2 for testing the null hypothesis of no primary effect.

References

Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociol. Methods Res.* **41**, 3–16.

Aronow, P. M. & Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Statist.* **11**, 1912–47.

Athey, S., Eckles, D. & Imbens, G. W. (2018). Exact p-values for network interference. *J. Am. Statist. Assoc.* **113**, 230–40.

Basse, G. & Feller, A. (2018). Analyzing two-stage experiments in the presence of interference. *J. Am. Statist. Assoc.* **113**, 41–55.

Bowers, J., Fredrickson, M. M. & Panagopoulos, C. (2013). Reasoning about interference between units: A general framework. *Polit. Anal.* **21**, 97–124.

Cox, D. R. (1958). *Planning of Experiments*. New York: John Wiley & Sons.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Geyer, C. J. & Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and p-values. *Statist. Sci.* **20**, 358–66.

Hennessy, J., Dasgupta, T., Miratrix, L., Pattanayak, C. & Sarkar, P. (2016). A conditional randomization test to account for covariate imbalance in randomized experiments. *J. Causal Infer.* **4**, 61–80.

Hudgens, M. G. & Halloran, M. E. (2008). Toward causal inference with interference. *J. Am. Statist. Assoc.* **103**, 832–42.

Manski, C. F. (2013). Identification of treatment response with social interactions. *Economet. J.* **16**, S1–S23.

Rigdon, J. & Hudgens, M. G. (2015). Exact confidence intervals in the presence of interference. *Statist. Prob. Lett.* **105**, 130–5.

Rogers, T. & Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Hum. Behav.* **2**, 335–42.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17**, 286–327.

Rubin, D. B. (1980). Comment. *J. Am. Statist. Assoc.* **75**, 591–3.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *J. Am. Statist. Assoc.* **101**, 1398–407.

Toulis, P. & Kao, E. (2013). Estimation of causal peer influence effects. *J. Mach. Learn. Res.* **28**, 1489–97.

[*Received on* 29 *July* 2017. *Editorial decision on* 20 *August* 2018]