The Annals of Statistics
2021, Vol. 49, No. 2, 673–701
https://doi.org/10.1214/20-AOS1973

AVERAGE TREATMENT EFFECTS IN THE PRESENCE OF UNKNOWN INTERFERENCE

BY FREDRIK SÄVJE[1], PETER M. ARONOW[2] AND MICHAEL G. HUDGENS[3]

[1]*Department of Political Science, Department of Statistics & Data Science, Yale University, fredrik.savje@yale.edu*
[2]*Department of Political Science, Department of Public Health (Biostatistics), Department of Statistics & Data Science, Yale University, peter.aronow@yale.edu*
[3]*Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, mhudgens@bios.unc.edu*

We investigate large-sample properties of treatment effect estimators under unknown interference in randomized experiments. The inferential target is a generalization of the average treatment effect estimand that marginalizes over potential spillover effects. We show that estimators commonly used to estimate treatment effects under no interference are consistent for the generalized estimand for several common experimental designs under limited but otherwise arbitrary and unknown interference. The rates of convergence depend on the growth rate of the unit-average amount of interference and the degree to which the interference aligns with dependencies in treatment assignment. Importantly for practitioners, the results imply that even if one erroneously assumes that units do not interfere in a setting with moderate interference, standard estimators are nevertheless likely to be close to an average treatment effect if the sample is sufficiently large. Conventional confidence statements may, however, not be accurate.

1. Introduction. Investigators of causality routinely assume that subjects under study do not interfere with each other. The no-interference assumption is so ingrained in the practice of causal inference that its use is often left implicit. Yet, interference is at the heart of the social and medical sciences. Humans interact, and these interactions are precisely the motivation for much of the research in these fields.

We investigate to what extent one can weaken the assumption of no interference and still draw credible inferences about causal relationships. We find that causal inference is impossible under completely unrestricted interference, so some assumptions must be made, but the conventional no-interference assumption is stronger than necessary. One can allow for moderate amounts of interference, and one can allow the subjects to interfere in unknown and largely arbitrary ways.

Our focus is the estimation of average treatment effects in randomized experiments. A random subset of a sample of units is assigned to some treatment, and the quantity of interest is the average effect of the assignment. The no-interference assumption in this context is the assertion that a unit's treatment assignment does not affect the outcome of any other unit. We consider the setting where such *spillover effects* do exist, and in particular, when the form they may take is left unspecified.

The paper makes four main contributions. We first introduce an estimand—the *expected average treatment effect* or EATE—that generalizes the conventional *average treatment effect* (ATE) to settings with interference. The conventional estimand is not well-defined when units interfere because the outcome of a unit may then be affected by more than one treatment.

Received October 2019; revised April 2020.
MSC2020 subject classifications. Primary 62G20; secondary 62D99, 62K99.
Key words and phrases. Causal effects, causal inference, experiments, SUTVA.

We resolve the issue by marginalizing the effects of interest over the assignment distribution of the incidental treatments. That is, for a given reference assignment, we ask how a particular unit's outcome is affected when we only change its own treatment assignment. An unambiguous average treatment effect is defined by asking the same for each unit in the experiment and averaging the resulting unit-level effects. While this average effect is unambiguous, it depends on which assignment was used as reference, and repeating the exercise with a different reference assignment can result in a different value. To capture the typical treatment effect in the experiment, EATE marginalizes these average effects over all possible reference assignments. The estimand is a generalization of ATE in the sense that they coincide whenever the latter is well-defined.

The second contribution is to demonstrate that EATE can be consistently estimated under weak restrictions on the interference and without structural knowledge thereof. We focus on the standard Horvitz–Thompson and Hájek estimators. The analysis also applies to the difference-in-means and ordinary least squares estimators, as they are special cases of the Hájek estimator. We begin by investigating the Bernoulli and complete randomization experimental designs. The estimators are consistent for EATE under these designs as long as the average amount of interference grows sufficiently slowly (according to measures we define shortly). Root-n consistency is achieved whenever the average amount of interference is bounded. We next investigate the paired randomization design. Paired randomization introduces perfectly correlated treatment assignments, and we show that this can make the estimators unstable even when the interference is limited. To achieve consistency, the degree to which the dependencies introduced by the experimental design align with the interference structure must be restricted. However, information about the interference structure beyond the aggregated restrictions is still not needed. We show that the insights from the paired design extend to a more general setting, and similar restrictions yield consistency under arbitrary experimental designs.

The third contribution is an investigation of variance estimation. We show that conventional variance estimators generally fail to capture the loss of precision that can result from interference. Confidence statements based on these estimators may therefore be misleading. To address the concern, we construct three alternative estimators by inflating a conventional estimator with measures of the amount of interference similar to those used to show consistency. Two of the alternative estimators are shown to be asymptotically conservative under weak conditions.

The final contribution is an investigation of whether EATE under one design generalizes to other designs. Because the estimand marginalizes over the design actually used in the experiment, it may have taken a different value if another design were used. We show that EATE for a given experiment is informative of the effect under designs that are close to the one that was implemented under suitable regularity conditions. When the amount of interference is limited, the estimands may converge.

The findings in this paper are of theoretical interest because they extend the known limits of causal inference under interference. They are also of practical interest. We investigate standard estimators under standard experimental designs, so the findings apply to many previous studies where interference might have been present but was assumed not to be. Therefore, studies that mistakenly assume that units do not interfere might not necessarily be invalidated. For example, no-interference assumptions are common in experimental studies of voter mobilization (see Green and Gerber (2004) and the references therein). However, a growing body of evidence suggests that potential voters interact within households, neighborhoods and other social structures (Aronow (2012), Nickerson (2008), Sinclair, McConnell and Green (2012)). Interference is, in other words, likely present in this setting, and researchers have been left uncertain about the interpretation of existing findings. Our results provide a lens through which the evidence can be interpreted; the reported estimates capture expected average treatment effects.

2. Related work. Our investigation builds on a recent literature on causal inference under interference (see Halloran and Hudgens (2016) for a review). The no-interference assumption was pervasive but implicit in the early literature on causal inference. The first explicit discussion of the assumption appears to have been by Cox (1958). The instantiation that is most commonly used today was formulated by Rubin (1980) as a part of the *stable unit treatment variation assumption*, or SUTVA.

Early departures from the no-interference assumption were modes of analysis inspired by Fisher's exact randomization test (Fisher (1935)). The approach uses sharp null hypotheses that stipulate the outcomes of all units under all assignments. The most common such hypothesis is that treatment is inconsequential, meaning that the observed outcomes would have been the same for all assignments. This subsumes the hypotheses that both primary and spillover effects do not exist, so the approach tests for the existence of both types of effects simultaneously. The test has been adapted and extended to study interference specifically (see, e.g., Aronow (2012), Athey, Eckles and Imbens (2018), Basse, Feller and Toulis (2019), Bowers, Fredrickson and Panagopoulos (2013), Choi (2017), Luo et al. (2012), Rosenbaum (2007)).

Early methods for point estimation restricted the interference process through structural models and thereby presumed that interactions were governed by a particular functional form (Manski (1993)). The structural approach has been extended to capture effects under weaker assumptions in a larger class of interference processes (Bramoullé, Djebbari and Fortin (2009), Graham (2008), Lee (2007)). Still, the approach has been criticized for being too restrictive (Angrist (2014), Goldsmith-Pinkham and Imbens (2013)).

A strand of the literature closer to the current study relaxes these structural assumptions. Interference is allowed to take arbitrary forms as long as it is contained within known and disjoint groups of units. The assumption is known as *partial interference* (see, e.g., Basse and Feller (2018), Hudgens and Halloran (2008), Kang and Imbens (2016), Liu and Hudgens (2014), Liu, Hudgens and Becker-Dreps (2016), Rigdon and Hudgens (2015), Tchetgen Tchetgen and VanderWeele (2012)). While partial interference allows for some progress on its own, it is often coupled with *stratified interference*. The additional assumption stipulates that the only relevant aspect of the interference is the proportion of treated units within each group. The identities of the units are, in other words, inconsequential for the spillover effects. Much like the structural approach, stratified interference restricts the form the interference can take.

More recent contributions have focused on relaxing the partial interference assumption. Interference is not restricted to disjoint groups, and units are allowed to interfere along general structures such as social networks (see, e.g., Aronow and Samii (2017), Basse and Airoldi (2018a), Eckles, Karrer and Ugander (2016), Forastiere, Airoldi and Mealli (2017), Jagadeesan, Pillai and Volfovsky (2017), Manski (2013), Ogburn and VanderWeele (2017), Sussman and Airoldi (2017), Tchetgen Tchetgen, Fulcher and Shpitser (2019), Toulis and Kao (2013), Ugander et al. (2013)). This relaxation allows for interactions of quite general forms, but the suggested estimation methods require detailed knowledge of the interference structure.

Previous investigations under unknown interference have primarily focused on the expectation of various estimators. Sobel (2006) derives the expectation of an instrumental variables estimator used in housing mobility experiments under unknown interference, and he shows that it is a mixture of primary and spillover effects for compilers and noncompilers. Hudgens and Halloran (2008) derive similar results for the average distributional shift effect, which we discuss in Section 3.3. A study by Egami (2017) considers the bias of treatment effect estimators when the interference can be described by a set of networks. Egami's framework includes a stratified interference assumption, but it admits general forms of interference because the networks are allowed to be overlapping and partially unobserved.

Unlike the focus in this paper, previous investigations under unknown interference either do not discuss the precision and consistency of the investigated estimators, or they do so only after assuming that the interference structure is known. One exception is a study by Basse and Airoldi (2018b). The authors consider average treatment effects under arbitrary and unknown interference just as we do, but they focus on inference about the contrast between the average outcome when all units are treated and the average outcome when no unit is treated. As we discuss in Section 3.3, this estimand provides a different description of the causal setting than EATE. In contrast to the findings in this paper, Basse and Airoldi show that no consistent estimator exists for their estimand even when the interference structure is known.

3. Treatment effects under interference.

3.1. *Preliminaries.* Consider a sample of n units indexed by the set $\mathbf{U} = \{1, 2, \ldots, n\}$. An experimenter intervenes on the world in ways that potentially affect the units. The intervention is described by a n-dimensional binary vector $\mathbf{z} = (z_1, z_2, \ldots, z_n) \in \{0, 1\}^n$. A particular value of $\mathbf{z}$ could, for example, denote that some drug is given to a certain subset of the units in $\mathbf{U}$. We are particularly interested in how unit i is affected by the ith dimension of $\mathbf{z}$. For short, we refer to z_i as unit i's treatment.

The effects of different interventions are defined as comparisons between the outcomes they produce. Each unit has a function $y_i \colon \{0, 1\}^n \to \mathbb{R}$ denoting the observed outcome for the unit under a specific and potentially counterfactual intervention (Holland (1986), Splawa-Neyman (1990)). In particular, $y_i(\mathbf{z})$ is the response of i when the intervention is $\mathbf{z}$. We refer to the elements of the image of this function as *potential outcomes*. It will prove convenient to write the potential outcomes in a slightly different form. Let $\mathbf{z}_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$ denote the $(n-1)$-dimensional vector constructed by deleting the ith element from $\mathbf{z}$. The potential outcome $y_i(\mathbf{z})$ can then be written as $y_i(z_i; \mathbf{z}_{-i})$.

Throughout the paper, we assume that the potential outcomes are well-defined. The assumption implies that the manner in which the experimenter manipulates $\mathbf{z}$ is inconsequential; no matter how $\mathbf{z}$ came to take a particular value, the outcome is the same. Well-defined potential outcomes also imply that no physical law or other circumstances prohibit $\mathbf{z}$ from taking any value in $\{0, 1\}^n$. This ensures that the potential outcomes are, indeed, potential. However, the assumption does not restrict the way the experimenter chooses to intervene on the world, and some interventions may have no probability of being realized.

The experimenter sets $\mathbf{z}$ according to a random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$. The probability distribution of $\mathbf{Z}$ is the *design* of the experiment. The design is the sole source of randomness we will consider. Let Y_i denote the observed outcome of unit i. The observed outcome is a random variable connected to the experimental design through the potential outcomes: $Y_i = y_i(\mathbf{Z})$. As above, $\mathbf{Z}_{-i}$ denotes $\mathbf{Z}$ without its ith element, so $Y_i = y_i(Z_i; \mathbf{Z}_{-i})$.

3.2. *Expected average treatment effects.* It is conventional to assume that the potential outcomes are restricted so a unit's outcome is only affected by its own treatment. That is, for any two assignments $\mathbf{z}$ and $\mathbf{z}'$, if the treatment of a given unit is the same for both assignments, then the outcome for that unit is also the same. This no-interference assumption admits a definition of the treatment effect τ_i for unit i as the contrast between its potential outcomes when we change its treatment:

$$\tau_i = y_i(1; \mathbf{z}_{-i}) - y_i(0; \mathbf{z}_{-i}),$$

where $\mathbf{z}_{-i}$ is any element of $\{0, 1\}^{n-1}$. No interference means that the choice of $\mathbf{z}_{-i}$ is inconsequential for the values of $y_i(1; \mathbf{z}_{-i})$ and $y_i(0; \mathbf{z}_{-i})$. The variable can therefore be left free without ambiguity, and it is common to use $y_i(z)$ as a shorthand for $y_i(z; \mathbf{z}_{-i})$. Experimenters often summarize the distribution of the unit-level treatment effects in a sample with its average.

DEFINITION 1. Under no interference, the *average treatment effect* (ATE) is the average unit-level treatment effect:

$$\tau_{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \tau_i.$$

The definition requires the no-interference assumption. References to *the* effect of a unit's treatment become ambiguous when units interfere because τ_i will then vary under permutations of $\mathbf{z}_{-i}$. The ambiguity is contagious; the average treatment effect is also ill-defined without the no-interference assumption.

To unambiguously talk about treatment effects under interference, we redefine the unit-level effect for unit i as the contrast between its potential outcomes when we change its treatment while holding all other treatments fixed at a given assignment $\mathbf{z}_{-i}$. We call this quantity the *assignment-conditional unit-level treatment effect*:

$$\tau_i(\mathbf{z}_{-i}) = y_i(1; \mathbf{z}_{-i}) - y_i(0; \mathbf{z}_{-i}).$$

To the best of our knowledge, this type of unit-level effect was first discussed by Halloran and Struchiner (1995). The assignment-conditional effect differs from τ_i only in that the dependence on the treatments of other units is made explicit. The redefined effect acknowledges that a unit's treatment may affect its outcome differently depending on the treatments assigned to other units. This makes the unit-level effects unambiguous, and their average provides a version of the average treatment effect that remains well-defined under interference.

DEFINITION 2. An *assignment-conditional average treatment effect* is the average of the assignment-conditional unit-level treatment effects under a given assignment:

$$\tau_{\text{ATE}}(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} \tau_i(\mathbf{z}_{-i}).$$

The assignment-conditional effects are well-defined under interference, but they are unwieldy. An average effect exists for each assignment, so their numbers grow exponentially in the sample size. For this reason, experimenters may not find it useful to study these effects individually. Similar to how unit-level effects are aggregated to an average effect, we focus on a summary of the distribution of the assignment-conditional effects.

DEFINITION 3. The *expected average treatment effect* (EATE) is the expected assignment-conditional average treatment effect:

$$\tau_{\text{EATE}} = \mathrm{E}\big[\tau_{\text{ATE}}(\mathbf{Z})\big],$$

where the expectation is taken over the distribution of $\mathbf{Z}$ given by the experimental design.

The expected average treatment effect is a generalization of ATE in the sense that the two estimands coincide whenever the no-interference assumption holds. Under no interference, $\tau_{\text{ATE}}(\mathbf{z})$ does not depend on $\mathbf{z}$, so the marginalization is inconsequential. When units interfere, $\tau_{\text{ATE}}(\mathbf{z})$ does depend on $\mathbf{z}$. The random variable $\tau_{\text{ATE}}(\mathbf{Z})$ is drawn from the distribution of the average treatment effect under the design that was used in the experiment. The EATE estimand provides the best description of this distribution in a mean square sense.

3.3. *Related definitions.* The EATE estimand builds on previously proposed ideas. An estimand introduced by Hudgens and Halloran (2008) resolves the ambiguity of treatment effects under interference in a similar way as we do. They refer to their estimand as the average direct causal effect, but we use the name average distributional shift effect to highlight how it differs from ATE and EATE.

DEFINITION 4. The *average distributional shift effect* (ADSE) is the average difference between the conditional expected outcomes for the two treatment conditions:

$$\tau_{\mathrm{ADSE}} = \frac{1}{n}\sum_{i=1}^{n}\big(\mathrm{E}[Y_i \mid Z_i = 1] - \mathrm{E}[Y_i \mid Z_i = 0]\big).$$

The effect marginalizes the potential outcomes over the experimental design, just as EATE does. The estimands differ in which distributions they use for the marginalization. The expectation in EATE is over the unconditional assignment distribution, while ADSE marginalizes each potential outcome separately over different conditional distributions. The difference becomes clear when the estimands are written in similar forms:

$$\tau_{\mathrm{EATE}} = \frac{1}{n}\sum_{i=1}^{n}\big(\mathrm{E}\big[y_i(1;\mathbf{Z}_{-i})\big] - \mathrm{E}\big[y_i(0;\mathbf{Z}_{-i})\big]\big),$$

$$\tau_{\mathrm{ADSE}} = \frac{1}{n}\sum_{i=1}^{n}\big(\mathrm{E}\big[y_i(1;\mathbf{Z}_{-i}) \mid Z_i = 1\big] - \mathrm{E}\big[y_i(0;\mathbf{Z}_{-i}) \mid Z_i = 0\big]\big).$$

The two estimands provide different causal information. EATE captures the expected average effect of changing the treatment of a single unit in the current experiment. ADSE captures the expected average effect of changing from an experimental design for which we hold a unit's treatment fixed at $Z_i = 1$ to another design for which its treatment is fixed at $Z_i = 0$. That is, the estimand captures the compound effect of changing the treatment of a unit and simultaneously changing the experimental design. As a result, ADSE may be nonzero even if all unit-level effects are exactly zero. That is, we may have $\tau_{\mathrm{ADSE}} \neq 0$ when $\tau_i(\mathbf{z}_{-i}) = 0$ for all i and $\mathbf{z}_{-i}$. Eck, Morozova and Crawford (2018) use a similar argument to show that ADSE may not correspond to causal parameters capturing treatment effects in structural models.

VanderWeele and Tchetgen Tchetgen (2011) introduced a version of ADSE that removes the artifact of the original estimand by conditioning both terms with the same value for the treatment of unit i. Hence, the marginalization is over the same distribution for both terms in the treatment effect contrast. Their estimand is a conditional average of unit-level effects and, as such, it mixes aspects of EATE and ADSE.

An alternative way to define an average treatment effect under interference is as the contrast between the average outcome when all units are treated and the average outcome when no unit is treated: $n^{-1}\sum_{i=1}^{n}[y_i(\mathbf{1}) - y_i(\mathbf{0})]$ where $\mathbf{1}$ and $\mathbf{0}$ are the unit and zero vectors. This *all-or-nothing effect* coincides with the conventional ATE in Definition 1 (and thus also with EATE) whenever the no-interference assumption holds. However, the effect does not coincide with EATE under interference, and the estimands provide different descriptions of the causal setting. EATE captures the typical treatment effect in the experiment actually implemented, while the all-or-nothing effect captures the effect of giving treatment to everyone or to no one. The all-or-nothing effect may therefore capture both primary and spillover effects. As we noted in Section 2, no consistent estimator exists for the effect in the context considered in this paper (Basse and Airoldi (2018b)).

3.4. *Example*. To build understanding about the EATE estimand, consider the following hypothetical vaccination study as an example. Let z_i denote whether person i received the vaccine under study, and let the outcome be whether the person eventually becomes infected by the pathogen the vaccine aims to protect against. We have reason to believe there is interference in this setting because a vaccinated person is less likely to carry the pathogen, and thus makes it less likely that other people become infected.

There are two types of effects we can investigate in this setting. The first is the primary, biological effect of the vaccine itself, corresponding to ATE in a setting without interference. That is, the effect on a person's outcome when we change whether the person itself is vaccinated. The second type is the spillover effects. That is, the effect on a person's outcome when we change whether other people are vaccinated. The purpose of EATE is to isolate the first type of effect.

To see how EATE differs from ADSE, consider a sample consisting of spouses in two-person households, and an experimental design that assigns the vaccine to exactly one person in each household. In this setting, the ADSE estimand captures the difference in outcomes when, on the one hand, a person receives the vaccine and their spouse does not, and on the other hand, when the person does not receive the vaccine but their spouse does. Because the vaccination status of both spouses change in this contrast, the estimand depends on both the primary effect of the vaccine and possible spillover effects between spouses. That is, depending on how the outcome of a person is affected by the vaccination status of their spouse, ADSE would take different values. This is the artifact of the ADSE estimand mentioned in the previous subsection.

The EATE estimand in this setting captures the difference in outcomes when, on the one hand, a person receives the vaccine, and on the other hand, when the person does not receive the vaccine, holding the vaccination statuses of the spouse and all other people fixed. The status of the spouse could be fixed at either receiving or not receiving the vaccine, and EATE takes the expectation over these two possible states. By doing so, the estimand acknowledges that the direct effect of the vaccine could differ depending on the vaccination status of the spouse, but unlike ADSE, the value it takes is itself not affected by the spillover effect between the spouses. In this way, the EATE estimand isolates the primary effect of the vaccine in the implemented experiment.

Figure 1 illustrates the difference between the EATE estimand and the all-or-nothing effect. We here consider a family of experimental designs that can be indexed by the expected proportion of vaccinated people. Both the Bernoulli and complete randomization designs investigated in Section 5.2 are of this type. The horizontal axis of the figure denotes this expected proportion. The vertical axis is the infection rate, which is the outcome, and the graphs depict the two expected average potential outcomes under the design given by the horizontal axis. In particular, they depict the functions

$$\bar{y}_1(p) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\big[y_i(1; \mathbf{Z}_{-i})\big] \quad \text{and} \quad \bar{y}_0(p) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\big[y_i(0; \mathbf{Z}_{-i})\big],$$

where p denotes the expected proportion of vaccinated units. The expectations in the expressions implicitly depend on p because the family of designs is indexed by the parameter.

The functions give the proportion of people that would become infected with and without the vaccine for different vaccination rates. The function $\bar{y}_1(p)$ takes a lower value than $\bar{y}_0(p)$ for all vaccination rates p, which captures that people generally have a lower risk of getting infected when vaccinated compared to when not vaccinated. Both functions decrease in p, which indicates that people are less likely to become infected as the vaccination rate increases independently of their own vaccination statuses. One possible explanation for this is that the community in which the study is run increasingly develops herd immunity as the vaccination

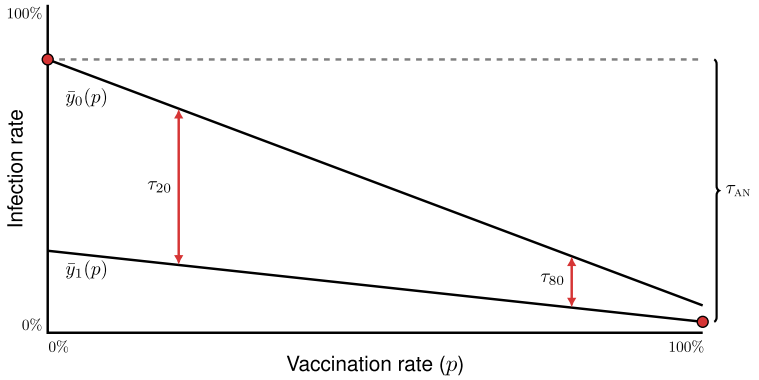

FIG. 1. *Illustration of two expected average treatment effects* (τ_{20} *and* τ_{80}) *and the all-or-nothing effect* (τ_{AN}).

rate increases, so all people get exposed to the pathogen to a lesser degree. However, the slope is steeper for the function $\bar{y}_0(p)$, demonstrating greater benefits of high vaccination rates for unvaccinated people. For very high vaccination rates, the difference between the functions is small. A possible explanation is that the pathogen is close to eradicated when almost everyone is vaccinated, so it is rare that people ever get exposed.

The EATE estimand captures the difference between $\bar{y}_1(p)$ and $\bar{y}_0(p)$ when p is the vaccination rate that was actually used in the experiment at hand. The figure depicts the estimand for two different experiments. The effect τ_{20} depicts EATE when the vaccination rate is 20%, and τ_{80} does the same when the rate is 80%. The difference between the two effects captures that the effect of the vaccine is larger when the vaccination rates are low. When $p = 20\%$, the EATE is negative and of large magnitude, indicating considerable reduction in the risk of becoming infected as a direct effect of the vaccine. When $p = 80\%$, the estimand is small, indicating only a slight reduction in the infection risk.

The all-or-nothing effect is the difference between the average outcome of vaccinated people when everyone is vaccinated, $\bar{y}_1(100)$, and the average outcome of unvaccinated people when no one is vaccinated, $\bar{y}_0(0)$. The figure depicts these two states with red-colored points, and the effect itself is depicted in the right-hand margin of the figure. Because the all-or-nothing effect simultaneously changes the vaccination status of everyone in the experiment, the effect captures both primary and spillover effects.

4. Quantifying interference. Our results do not require detailed structural information about the interference. However, as we show in Section 5.1, no progress can be made if it is left completely unrestricted. The following definitions quantify the amount of interference in an experiment and will serve as the basis for the restrictions we use in the paper.

We say that unit i interferes with unit j if changing i's treatment changes j's outcome under at least one treatment assignment. We also say that a unit interferes with itself even if its own treatment does not affect its outcome. The indicator I_{ij} denotes such interference:

$$I_{ij} = \begin{cases} 1 & \text{if } y_j(\mathbf{z}) \neq y_j(\mathbf{z}') \text{ for some } \mathbf{z}, \mathbf{z}' \in \{0,1\}^n \text{ such that } \mathbf{z}_{-i} = \mathbf{z}'_{-i}, \\ 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

The definition of I_{ij} allows for asymmetric interference in the sense that unit i may interfere with unit j even when the converse does not hold.

The collection of interference indicators only describes the interference structure in an experimental sample. The definition itself does not restrict how the units may interfere. In

particular, the indicators do not necessarily align with social networks or other structures through which units are thought to interact. Experimenters do not generally have enough information about how the units interfere to deduce or estimate the indicators. The role of the interference indicators is instead to act as the foundation of the following aggregated summary.

DEFINITION 5 (Interference dependence).

$$d_{\mathrm{AVG}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} \quad \text{where } d_{ij} = \begin{cases} 1 & \text{if } I_{\ell i} I_{\ell j} = 1 \text{ for some } \ell \in \mathbf{U}, \\ 0 & \text{otherwise.} \end{cases}$$

The interference dependence indicator d_{ij} captures whether units i and j are affected by a common treatment. That is, i and j are interference dependent if they interfere directly with each other or if some third unit interferes with both i and j. The sum $d_i = \sum_{j=1}^{n} d_{ij}$ gives the number of interference dependencies for unit i, so d_{AVG} is the unit-average number of interference dependencies.

The average interference dependence d_{AVG} quantifies how close an experiment is to no interference. Complete absence of interference is the same as $d_{\mathrm{AVG}} = 1$, which indicates that each unit is only interfering with itself. At the other extreme, $d_{\mathrm{AVG}} = n$ indicates that interference is complete in the sense that all pairs of units are affected by a common treatment. If sufficiently many units are interference dependent (i.e., if d_{AVG} is large), small perturbations of the treatment assignments may be amplified by the interference and induce large changes in the outcomes of many units.

Interference dependence can be related to simpler descriptions of the interference. Consider the following quantities:

$$c_i = \sum_{j=1}^{n} I_{ij} \quad \text{and} \quad C_p = \left[\frac{1}{n}\sum_{i=1}^{n} c_i^p\right]^{1/p}.$$

The first quantity c_i captures how many units i interferes with. That is, if changing the treatment of unit i would change the outcome of five other units, then unit i interferes with those five units and itself, so $c_i = 6$. Information about these quantities are generally beyond the grasp of experimenters. The quantity C_p provides a more aggregated description, which experimenters may have more insights about. This is the p-norm of the unit-level interference counts with respect to the sample measure. For example, C_1 and C_2 are the average and root mean square of the unit-level quantities. We write C_∞ for the limit of C_p as $p \to \infty$, which is the maximum c_i over $\mathbf{U}$. These norms bound d_{AVG} from below and above.

LEMMA 1. $\max(C_1, n^{-1}C_\infty^2) \le d_{\mathrm{AVG}} \le C_2^2 \le C_\infty^2$.

All proofs, including that of Lemma 1, are given in Supplement A (Sävje, Aronow and Hudgens (2021)). The lemma implies that we can use C_2 or C_∞, rather than d_{AVG}, to restrict the interference. While such restrictions are stronger than necessary, the connection is nevertheless useful because experimenters may find it more intuitive to reason about the norms of c_i than about interference dependence.

5. Large sample properties. Inspired by an asymptotic regime in Isaki and Fuller (1982), we consider an arbitrary sequence of samples indexed by their sample size. The samples are not assumed to be drawn from some larger population or otherwise randomly generated. That is, the samples are not necessarily related other than through the conditions

we impose on the sequences. All quantities describing the samples, such as the potential outcomes and experimental designs, have their own sequences also indexed by n. But we leave the indexing implicit as no confusion arises. The focus of the investigation is how two estimators of average treatment effects behave as the sample size grows.

DEFINITION 6. The Horvitz–Thompson (HT) and Hájek (HÁ) estimators are

$$\hat{\tau}_{\mathrm{HT}} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{p_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-p_i} \quad \text{and}$$

$$\hat{\tau}_{\mathrm{HÁ}} = \left(\sum_{i=1}^{n}\frac{Z_i Y_i}{p_i}\Big/\sum_{i=1}^{n}\frac{Z_i}{p_i}\right) - \left(\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-p_i}\Big/\sum_{i=1}^{n}\frac{1-Z_i}{1-p_i}\right),$$

where $p_i = \Pr(Z_i = 1)$ is the marginal treatment probability for unit i.

Estimators of this form were first introduced in the sampling literature to estimate population means under unequal inclusion probabilities (Hájek (1971), Horvitz and Thompson (1952)). They have since received much attention in the causal inference and policy evaluation literatures where they are often called *inverse probability weighted estimators* (see, e.g., Hahn (1998), Hernán and Robins (2006), Hirano, Imbens and Ridder (2003)). Other estimators commonly used to analyze experiments, such as the difference-in-means and ordinary least squares estimators, are special cases of the Hájek estimator. Consequently, our analysis applies to those estimators as well.

We assume throughout the remainder of the paper that the experimental design and potential outcomes are sufficiently well behaved, as formalized in the following assumption.

ASSUMPTION 1 (Regularity conditions). There exist constants $k < \infty$, $q \geq 2$ and $s \geq 1$ such that for all $i \in \mathbf{U}$ in the sequence of samples:

A (Probabilistic assignment). $k^{-1} \leq \Pr(Z_i = 1) \leq 1 - k^{-1}$,
B (Outcome moments). $\mathrm{E}[|Y_i|^q] \leq k^q$,
C (Potential outcome moments). $\mathrm{E}[|y_i(z; \mathbf{Z}_{-i})|^s] \leq k^s$ for $z \in \{0, 1\}$.

The first regularity condition restricts the experimental design so that each treatment is realized with a positive probability. The condition does not restrict combinations of treatments, and some assignment vectors may have no probability of being realized. The second condition restricts the distributions of the observed outcomes so they remain well behaved asymptotically. The last condition restricts the potential outcomes slightly off the support of the experimental design and ensures that EATE is well-defined asymptotically.

The exact values of q and s are inconsequential for the results in Section 5.2. The weakest version of the assumption, when $q = 2$ and $s = 1$, suffices there. However, the rate of convergence under an arbitrary experimental design depends on which moments exist, and variance estimation generally requires that q is at least four. The ideal case is when the potential outcomes themselves are asymptotically bounded, which corresponds to the case where Assumption 1 holds as $q \to \infty$ and $s \to \infty$.

The two moment conditions are similar in structure, but neither is implied by the other. Assumption 1B does not imply Assumption 1C because the former is only concerned with the potential outcomes on the support of the experimental design. The opposite implication does not hold because s may be smaller than q.

5.1. *Restricting the interference.* The sequence of d_{AVG} describes the amount of interference in the sequence of samples. Our notion of limited interference is formalized as a restriction on this sequence.

ASSUMPTION 2 (Restricted interference). $d_{\text{AVG}} = \text{o}(n)$.

The assumption stipulates that units, on average, are interference dependent with an asymptotically diminishing fraction of the sample. The assumption still allows for substantial amounts of interference. The unit-average number of interference dependencies may, for example, grow with the sample size and the total number of interference dependencies may grow at a faster rate than n. What is assumed is that the unit-average d_{AVG} does not grow proportionally to the sample size.

In addition to restricting the amount of interference, Assumption 2 imposes weak restrictions on the structure of the interference. It prohibits interference that is so unevenly distributed that a few units are interfering with most other units. If the interference is concentrated in such a way, small perturbations of the assignments could be amplified through the treatments of those units. At the extreme when a single unit interferes with all other units, the outcomes of all units would change if we were to change the treatment of that single unit. The estimators would then not stabilize even if the interference was otherwise sparse.

Restricted interference is not sufficient for consistency. Sequences of experiments exist for which Assumption 2 holds but the estimators do not converge to EATE. However, the assumption is necessary for consistency of the HT and HÁ estimators in the following sense.

PROPOSITION 1. *For every sequence of experimental designs, if Assumption 2 does not hold, there exists a sequence of potential outcomes satisfying Assumption 1 such that the* HT *and* HÁ *estimators do not converge in probability to* EATE.

The proposition implies that the weakest possible restriction on d_{AVG} is Assumption 2. If a weaker restriction is imposed, for example, that d_{AVG} is on the order of εn for some small $\varepsilon > 0$, then there exist potential outcomes for any experimental design such that the relaxed interference restriction is satisfied but the estimators do not converge. A consequence is that experimental designs themselves cannot ensure consistency. We must somehow restrict the interference to make progress, and in this sense, our restricted interference assumption is necessary for our results. It might, however, be possible to achieve consistency without Assumption 2 if one were to impose stronger regularity conditions or restrict the interference in some other way. For example, the estimators could be consistent if the magnitude of the interference, according to some suitable measure, approaches zero.

5.2. *Common experimental designs.* Our large sample investigation begins with three specific experimental designs. These designs are commonly used by experimenters, and they are therefore of interest in their own right. They also provide a good illustration of the issues that arise under unknown interference, setting the scene for the investigation of arbitrary designs in the Section 5.3.

5.2.1. *Bernoulli and complete randomization.* The simplest experimental design assigns the treatments independently. The experimenter flips a coin for each unit and administers treatment accordingly. We call this a *Bernoulli randomization design*, and it satisfies

$$\Pr(\mathbf{Z} = \mathbf{z}) = \prod_{i=1}^{n} p_i^{z_i} (1 - p_i)^{1 - z_i}$$

for some set of assignment probabilities $p_1, p_2, \ldots, p_n$ bounded away from zero and one.

The outcomes of any two units are independent under a Bernoulli design when the no-interference assumption holds. This is not the case when units interfere. A single treatment may then affect two or more units, and the corresponding outcomes are dependent. That is, two units' outcomes are dependent when they are interference dependent according to Definition 5. Restricting this dependence ensures that the effective sample size grows with the nominal size and yields consistency.

PROPOSITION 2. *With a Bernoulli randomization design under restricted interference (Assumption 2), the* HT *and* HÁ *estimators are consistent for* EATE *and converge at the following rates*:

$$\hat{\tau}_{\mathrm{HT}} - \tau_{\mathrm{EATE}} = \mathcal{O}_{\mathrm{p}}(n^{-0.5} d_{\mathrm{AVG}}^{0.5}) \quad \textit{and} \quad \hat{\tau}_{\mathrm{HÁ}} - \tau_{\mathrm{EATE}} = \mathcal{O}_{\mathrm{p}}(n^{-0.5} d_{\mathrm{AVG}}^{0.5}).$$

The Bernoulli design tends to be inefficient in small samples because the size of the treatment group varies over assignments. Experimenters often prefer designs that reduce the variability in the group sizes. One common such design randomly selects an assignment with equal probability from all assignments with a certain proportion of treated units:

$$\Pr(\mathbf{Z} = \mathbf{z}) = \begin{cases} \binom{n}{m}^{-1} & \text{if } \sum_{i=1}^{n} z_i = m, \\ 0 & \text{otherwise,} \end{cases}$$

where $m = \lfloor pn \rfloor$ for some fixed p strictly between zero and one. The parameter p controls the desired proportion of treated units. We call the design *complete randomization*.

Complete randomization introduces dependencies between assignments. These dependencies are not of concern when there is no interference. The outcomes are only affected by a single treatment, and the dependence between any two treatments is asymptotically negligible. This need not be the case when units interfere; there are two issues to consider.

The first issue is that the interference could interact with the experimental design so that two units' outcomes are strongly dependent asymptotically even when they are not affected by a common treatment (i.e., when $d_{ij} = 0$). As an example, consider when one unit is affected by the first half of the sample and another unit is affected by the second half. Complete randomization introduces a strong dependence between the two halves: the number of treated units in the first half is perfectly correlated with the number of treated units in the second half. The outcomes of the two units may therefore be (perfectly) correlated even when no treatment affects them both. We cannot rule out that such dependencies exist, but we can show that they are sufficiently rare to not prevent convergence under a slightly stronger version of Assumption 2.

The second issue is that the dependencies introduced by the design distort our view of the potential outcomes. Whenever a unit is assigned to a certain treatment condition, units that interfere with that unit tend to be assigned to the other condition. One of the potential outcomes in each assignment-conditional unit-level effect is therefore observed more frequently than the other. The estimators implicitly weight the two potential outcomes proportionally to their frequency, but the EATE estimand weights them equally. This discrepancy introduces bias. Seen from another perspective, the estimators do not separate the effect of a unit's own treatment from spillover effects of other units' treatments.

As an illustration, consider when the potential outcomes are equal to the number of treated units: $y_i(\mathbf{z}) = \sum_{j=1}^{n} z_j$. EATE equals one in this case, but the estimators are constant at zero because the number of treated units (and thus, all revealed potential outcomes) are fixed at m. The design exactly masks the effect of a unit's own treatment with a spillover effect of the same magnitude but of the opposite sign.

Under complete randomization, if the number of units interfering with a given unit is of the same order as the sample size, our view of the unit's potential outcomes will also be distorted

asymptotically. As with the first issue, we cannot rule out that such distortions exist, but restricted interference implies that they are sufficiently rare. Taken together, this establishes consistency under complete randomization.

PROPOSITION 3. *With a complete randomization design under restricted interference (Assumption 2) and* $C_1 = o(n^{0.5})$, *the* HT *and* HÁ *estimators are consistent for* EATE *and converge at the following rates*:

$$\hat{\tau}_{\text{HT}} - \tau_{\text{EATE}} = \mathcal{O}_p(n^{-0.5} d_{\text{AVG}}^{0.5} + n^{-0.5} C_1),$$
$$\hat{\tau}_{\text{HÁ}} - \tau_{\text{EATE}} = \mathcal{O}_p(n^{-0.5} d_{\text{AVG}}^{0.5} + n^{-0.5} C_1).$$

The proposition requires $C_1 = o(n^{0.5})$ in addition to Assumption 2. Both C_1^2 and d_{AVG} are bounded from above by C_2^2, so they tend to not be too different. It is when d_{ij} largely aligns with I_{ij} that C_1^2 dominates d_{AVG}. For example, we have $C_1 = d_{\text{AVG}}$ when all interference dependent units are interfering with each other directly, because then $I_{ij} = d_{ij}$.

The HT and HÁ estimators are known to be root-n consistent for ATE under no interference. Reassuringly, the no-interference assumption is equivalent to the condition $d_{\text{AVG}} = C_1 = 1$, and Propositions 2 and 3 provide root-n consistency. However, the propositions make clear that absence of interference is not necessary for such convergence rates, and we may allow for nontrivial amounts of interference. In particular, the estimators are root-n consistent whenever the interference dependence does not grow indefinitely with the sample size, that is, when d_{AVG} is bounded.

COROLLARY 1. *With a Bernoulli or complete randomization design under bounded interference,* $d_{\text{AVG}} = \mathcal{O}(1)$, *the* HT *and* HÁ *estimators are root-n consistent for* EATE.

5.2.2. *Paired randomization.* Complete randomization restricts the assignment of treatments to ensure treatment groups of fixed size. The paired randomization design imposes even greater restrictions. The sample is divided into pairs, and the units in each pair are assigned to different treatments. It is implicit that the sample size is even so that all units are paired. Paired randomization could be forced on the experimenter by external constraints or used to improve precision (see, e.g., Fogarty (2018) and the references therein).

Let $\rho : \mathbf{U} \to \mathbf{U}$ describe a pairing so that $\rho(i) = j$ indicates that units i and j are paired. The pairing is symmetric, so the self-composition of ρ is the identity function. The *paired randomization design* then satisfies

$$\Pr(\mathbf{Z} = \mathbf{z}) = \begin{cases} 2^{-n/2} & \text{if } z_i \neq z_{\rho(i)} \text{ for all } i \in \mathbf{U}, \\ 0 & \text{otherwise.} \end{cases}$$

The design worsens both issues we faced under complete randomization. Under paired randomization, Z_i and Z_j are perfectly correlated, also asymptotically, whenever $\rho(i) = j$. We must therefore consider to what extent the dependencies between assignments introduced by the design align with the structure of the interference. The following two definitions quantify the alignment.

DEFINITION 7 (Pair-induced interference dependence).

$$e_{\text{AVG}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} e_{ij} \quad \text{where } e_{ij} = \begin{cases} 1 & \text{if } (1 - d_{ij}) I_{\ell i} I_{\rho(\ell) j} = 1 \text{ for some } \ell \in \mathbf{U}, \\ 0 & \text{otherwise.} \end{cases}$$

DEFINITION 8 (Within-pair interference). $R_{\text{SUM}} = \sum_{i=1}^{n} I_{\rho(i) i}$.

The dependence within any set of finite number of treatments is asymptotically negligible under complete randomization, and issues only arose when the number of treatments affecting a unit was of the same order as the sample size. Under paired randomization, the dependence between the outcomes of two units not affected by a common treatment can be asymptotically nonnegligible even when each unit is affected by an asymptotically negligible fraction of the sample. In particular, the outcomes of units i and j such that $d_{ij} = 0$ can be (perfectly) correlated if two other units k and ℓ exist such that k interferes with i and ℓ interferes with j, and k and ℓ are paired. The purpose of Definition 7 is to capture such dependencies. The definition is similar in structure to Definition 5. Indeed, the upper bound from Lemma 1 applies so that $e_{\text{AVG}} \leq C_2^2$.

The second issue we faced under complete randomization is also made worse under paired randomization. No matter the number of units that are interfering with unit i, if one of those units is the unit paired with i, we cannot separate the effects of Z_i and $Z_{\rho(i)}$. The design imposes $Z_i = 1 - Z_{\rho(i)}$, so any effect of Z_i on i's outcome could just as well be attributed to $Z_{\rho(i)}$. Such dependencies introduce bias, just as they did under complete randomization. However, unlike complete randomization, restricted interference does not imply that the bias will vanish as the sample grows. We must separately ensure that this type of alignment between the design and the interference is sufficiently rare. The purpose of Definition 8 is to captures how common interference is between paired units.

The two definitions allow us to restrict the degree to which the interference aligns with the pairing in the design.

ASSUMPTION 3 (Restricted pair-induced interference). $e_{\text{AVG}} = \text{o}(n)$.

ASSUMPTION 4 (Pair separation). $R_{\text{SUM}} = \text{o}(n)$.

Experimenters may find that Assumption 3 is quite tenable under restricted interference. As both e_{AVG} and d_{AVG} are bounded by C_2^2, restricted pair-induced interference tends to hold in cases where restricted interference can safely be assumed. It is, however, possible that the latter assumption holds even when the former does not if paired units are interfering with sufficiently disjoint sets of units.

Whether pair separation holds depends largely on how the pairs were formed. It is, for example, common that the pairs reflect some social structure Paired units may, for example, live in the same household. The interference tends to align with the pairing in such cases, and Assumption 4 is unlikely to hold. Pair separation is more reasonable when pairs are formed based on generic background characteristics. This is often the case when the experimenter uses the design to increase precision. The assumption could, however, still be violated if the background characteristics include detailed geographic data or other information likely to be associated with the interference.

PROPOSITION 4. *With a paired randomization design under restricted interference, restricted pair-induced interference and pair separation* (*Assumptions* 2, 3 *and* 4), *the* HT *and* HÁ *estimators are consistent for* EATE *and converge at the following rates*:

$$\hat{\tau}_{\text{HT}} - \tau_{\text{EATE}} = \mathcal{O}_{\text{p}}(n^{-0.5} d_{\text{AVG}}^{0.5} + n^{-0.5} e_{\text{AVG}}^{0.5} + n^{-1} R_{\text{SUM}}),$$
$$\hat{\tau}_{\text{HÁ}} - \tau_{\text{EATE}} = \mathcal{O}_{\text{p}}(n^{-0.5} d_{\text{AVG}}^{0.5} + n^{-0.5} e_{\text{AVG}}^{0.5} + n^{-1} R_{\text{SUM}}).$$

5.3. *Arbitrary experimental designs*. We conclude this section by considering sequences of experiments with unspecified designs. Arbitrary experimental designs may align with the

interference just like the paired design. We begin by introducing a set of definitions that allow us to characterize such alignment in a general setting.

It will prove useful to collect all treatments affecting a particular unit i into a vector:

$$\widetilde{\mathbf{Z}}_i = (I_{1i}Z_1, I_{2i}Z_2, \ldots, I_{ni}Z_n).$$

The vector is defined so that its jth element is Z_j if unit j is interfering with i, and zero otherwise. Let $\widetilde{\mathbf{Z}}_{-i}$ be the $(n-1)$-dimensional vector constructed by deleting the ith element from $\widetilde{\mathbf{Z}}_i$. The definitions have the following convenient properties:

$$Y_i = y_i(\mathbf{Z}) = y_i(\widetilde{\mathbf{Z}}_i) \quad \text{and} \quad y_i(z; \mathbf{Z}_{-i}) = y_i(z; \widetilde{\mathbf{Z}}_{-i}).$$

We can characterize the outcome dependence introduced by the experimental design by the dependence between $\widetilde{\mathbf{Z}}_i$ and $\widetilde{\mathbf{Z}}_j$. Because $Y_i = y_i(\widetilde{\mathbf{Z}}_i)$, the outcomes of two units i and j are independent whenever $\widetilde{\mathbf{Z}}_i$ and $\widetilde{\mathbf{Z}}_j$ are independent. Similarly, the dependence between Z_i and $\widetilde{\mathbf{Z}}_{-i}$ governs how distorted our view of the potential outcomes is.

We use the alpha-mixing coefficient introduced by Rosenblatt (1956) to measure the dependence between the assignment vectors. Specifically, for two random variables X and Y defined on the same probability space, let

$$\alpha(X, Y) = \sup_{\substack{x \in \sigma(X) \\ y \in \sigma(Y)}} \big|\Pr(x \cap y) - \Pr(x)\Pr(y)\big|,$$

where $\sigma(X)$ and $\sigma(Y)$ denote the sub-sigma-algebras generated by the random variables. The coefficient $\alpha(X, Y)$ is zero if and only if X and Y are independent, and increasing values indicate increasing dependence. The maximum is $\alpha(X, Y) = 1/4$. Unlike the Pearson correlation coefficient, the alpha-mixing coefficient is not restricted to linear associations between two scalar random variables, and it can capture any type of dependence between any two sets of random variables. The coefficient allows us to define measures of the average amount of dependence between $\widetilde{\mathbf{Z}}_i$ and $\widetilde{\mathbf{Z}}_j$ and between Z_i and $\widetilde{\mathbf{Z}}_{-i}$.

DEFINITION 9 (External and internal average mixing coefficients). For the maximum values of q and s such that Assumptions 1B and 1C hold, let

$$\alpha_{\text{EXT}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(1 - d_{ij})\big[\alpha(\widetilde{\mathbf{Z}}_i, \widetilde{\mathbf{Z}}_j)\big]^{\frac{q-2}{q}} \quad \text{and} \quad \alpha_{\text{INT}} = \sum_{i=1}^{n}\big[\alpha(Z_i, \widetilde{\mathbf{Z}}_{-i})\big]^{\frac{s-1}{s}},$$

where 0^0 is defined as zero to accommodate the cases $q = 2$ and $s = 1$.

The terms of the external mixing coefficient capture the dependence between the treatments affecting unit i and the treatments affecting unit j. If the dependence between $\widetilde{\mathbf{Z}}_i$ and $\widetilde{\mathbf{Z}}_j$ tends to be weak or rare, α_{EXT} will be small compared to n. Similarly, if dependence between Z_i and $\widetilde{\mathbf{Z}}_{-i}$ tends to be weak or rare, α_{INT} will be small relative to n. In this sense, the external and internal mixing coefficients are generalizations of Definitions 7 and 8. Indeed, one can show that $\alpha_{\text{EXT}} \propto e_{\text{AVG}}$ and $\alpha_{\text{INT}} \propto R_{\text{SUM}}$ under paired randomization, where the proportionality constants are given by q and s.

The generalized definitions allow for generalized assumptions.

ASSUMPTION 5 (Design mixing). $\alpha_{\text{EXT}} = \mathrm{o}(n)$.

ASSUMPTION 6 (Design separation). $\alpha_{\text{INT}} = \mathrm{o}(n)$.

Design mixing and design separation stipulate that the dependence between treatments are sufficiently rare or sufficiently weak (or some combination thereof). The assumptions encapsulate and extend the conditions in the previous sections. In particular, complete randomization under bounded interference constitutes a setting where dependence is weak: $\alpha(\widetilde{\mathbf{Z}}_i, \widetilde{\mathbf{Z}}_j)$ approaches zero for all pairs of units with $d_{ij} = 0$. Paired randomization under Assumption 3 constitutes a setting where dependence is rare: $\alpha(\widetilde{\mathbf{Z}}_i, \widetilde{\mathbf{Z}}_j)$ may be $1/4$ for some pairs of units with $d_{ij} = 0$, but such pairs are an asymptotically diminishing fraction of the total number of pairs. Complete randomization under the conditions of Proposition 3 combines the two settings: $\alpha(\widetilde{\mathbf{Z}}_i, \widetilde{\mathbf{Z}}_j)$ might be nonnegligible asymptotically for some pairs with $d_{ij} = 0$, but such pairs are rare. For all other pairs with $d_{ij} = 0$, the pair-level mixing coefficient quickly approaches zero. A similar comparison can be made for the design separation assumption.

PROPOSITION 5. *Under restricted interference, design mixing and design separation (Assumptions 2, 5 and 6), the* HT *and* HÁ *estimators are consistent for* EATE *and converge at the following rates*:

$$\hat{\tau}_{\mathrm{HT}} - \tau_{\mathrm{EATE}} = \mathcal{O}_{\mathrm{p}}\big(n^{-0.5} d_{\mathrm{AVG}}^{0.5} + n^{-0.5} \alpha_{\mathrm{EXT}}^{0.5} + n^{-1} \alpha_{\mathrm{INT}}\big),$$
$$\hat{\tau}_{\mathrm{HÁ}} - \tau_{\mathrm{EATE}} = \mathcal{O}_{\mathrm{p}}\big(n^{-0.5} d_{\mathrm{AVG}}^{0.5} + n^{-0.5} \alpha_{\mathrm{EXT}}^{0.5} + n^{-1} \alpha_{\mathrm{INT}}\big).$$

REMARK 1. The convergence results for Bernoulli and paired randomization presented in the previous subsections can be proven as consequences of Proposition 5. This is not the case for complete randomization. The current proposition applied to that design would suggest slower rates of convergence than given by Proposition 3. This highlights that Proposition 5 provides worst-case rates for all designs that satisfy the stated conditions. Particular designs might be better behaved and thus ensure that the estimators converge at faster rates. For complete randomization, one can prove that restricted interference implies a mixing condition that is stronger than the conditions defined above. In particular, Lemmas A12 and A13 in Supplement A (Sävje, Aronow and Hudgens (2021)) provide bounds on the external and internal mixing coefficients when they are redefined using the mixing concept similar to the one introduced by Blum, Hanson and Koopmans (1963). Proposition 3 follows from this stronger mixing property.

REMARK 2. If no units interfere, $\widetilde{\mathbf{Z}}_{-i}$ is constant at zero, and Assumption 6 is trivially satisfied. However, no interference does not imply that Assumption 5 holds. Consider a design that assigns the same treatment to all units: $Z_1 = Z_2 = \cdots = Z_n$. The external mixing coefficient would not be zero in this case; in fact, $\alpha_{\mathrm{EXT}} \to n/4$. This example illustrates that one must limit the dependencies between treatment assignments even when there is no interference. Proposition 5 can, in this sense, be seen as an extension of Theorem 1 in Robinson (1982).

5.4. *When design separation fails*. Experimental designs tend to induce dependence between treatments of units that interfere with one another, and experimenters might find it hard to satisfy design separation. We saw one example of such a design with paired randomization. It might for this reason be advisable to choose uniform designs such as the Bernoulli or complete randomization when one investigates treatment effects under unknown interference. These designs cannot align with the interference structure, and one need only consider whether the simpler interference conditions hold. Another approach is to design the experiment in a way that ensures design separation. For example, one should avoid pairing units that are suspected to interfere in the paired randomization design.

However, it will not always be possible to ensure that design separation holds. The question is then what the consequences of such departures are. Without Assumption 6, the effect of a unit's own treatment on its outcome cannot be separated from potential spillover effects, and the estimators need not be consistent for EATE. But, they do converge to another quantity. Recall from Definition 4 that the average distributional shift effect uses the conditional distributions of the outcomes. As a consequence, the estimand does not attempt to completely separate the effect of a unit's own treatment from spillover effects, and design separation is not needed for consistency.

PROPOSITION 6. *Under restricted interference and design mixing* (*Assumptions 2 and 5*), *the* HT *and* HÁ *estimators are consistent for* ADSE *and converge at the following rates*:

$$\hat{\tau}_{\mathrm{HT}} - \tau_{\mathrm{ADSE}} = \mathcal{O}_{\mathrm{p}}(n^{-0.5} d_{\mathrm{AVG}}^{0.5} + n^{-0.5} \alpha_{\mathrm{EXT}}^{0.5}),$$
$$\hat{\tau}_{\mathrm{HÁ}} - \tau_{\mathrm{ADSE}} = \mathcal{O}_{\mathrm{p}}(n^{-0.5} d_{\mathrm{AVG}}^{0.5} + n^{-0.5} \alpha_{\mathrm{EXT}}^{0.5}).$$

The proposition highlights the connection between the EATE and ADSE estimands. With the exception of design separation, Proposition 6 uses the same assumptions as Proposition 5. Hence, when the assumptions of Proposition 5 hold, the estimators are consistent for both EATE and ADSE, and the two estimands coincide asymptotically. To understand why this is, recall that EATE may depend on potential outcomes slightly outside the support of the design. The ADSE estimand is, however, defined to coincide with the expectation of the HT estimator, so it depends only on potential outcomes on the support. The purpose of design separation, and the corresponding assumptions in Section 5.2, is to allow us to do the small extrapolation needed to learn the potential outcomes used in the definition of EATE that are outside the support.

6. Confidence statements. Experimenters often present point estimates of treatment effects together with statements about the precision of the estimation method. These statements should be interpreted with caution in the presence of unknown interference, because the precision of the estimator may be worse when units interfere. This is clear from the rates of convergence presented in the previous section. Disregarding outlandish experimental designs, the estimators we investigate converge at a root-n rate under no interference, but the propositions in the previous section show that the estimators may converge at a slower rate when units interfere. The question in this section is whether we can construct variance estimators that accurately reflect this potential loss in precision.

6.1. *A conventional variance estimator.* We illustrate the issues that can arise under unknown interference by investigating the validity of a conventional variance estimator. To avoid some technical difficulties of little relevance to the current discussion, we restrict our focus to the Horvitz–Thompson estimator of the variance of the Horvitz–Thompson point estimator under a Bernoulli design:

$$\widehat{\mathrm{Var}}_{\mathrm{BER}}(\hat{\tau}_{\mathrm{HT}}) = \frac{1}{n^2} \sum_{i=1}^{n} \frac{Z_i Y_i^2}{p_i^2} + \frac{1}{n^2} \sum_{i=1}^{n} \frac{(1 - Z_i) Y_i^2}{(1 - p_i)^2}.$$

The estimator is conservative under no interference, meaning that its expectation is greater than the true variance. On a normalized scale, the bias does not diminish asymptotically, so inferences based on the estimator will be conservative also in large samples. In particular, with a Bernoulli design under no interference,

$$n\big[\widehat{\mathrm{Var}}_{\mathrm{BER}}(\hat{\tau}_{\mathrm{HT}}) - \mathrm{Var}(\hat{\tau}_{\mathrm{HT}})\big] \xrightarrow{p} \tau_{\mathrm{MSQ}} \geq 0, \tag{1}$$

where τ_{MSQ} is the mean square treatment effect:

$$\tau_{\text{MSQ}} = \frac{1}{n}\sum_{i=1}^{n}\big(\mathrm{E}\big[\tau_i(\mathbf{Z}_{-i})\big]\big)^2.$$

We define τ_{MSQ} using $\tau_i(\mathbf{Z}_{-i})$ rather than τ_i, because this will allow us to use the same definition both when interference is present and when it is not. In the current setting, we could have used τ_i because $\tau_i(\mathbf{z}_{-i})$ does not depend on $\mathbf{z}_{-i}$ when units do not interfere.

To characterize the behavior of the variance estimator under interference, it is helpful to introduce additional notation. Let $\xi_{ij}(z)$ be the expected treatment effect on unit i's outcome when changing unit j's treatment given that i is assigned to treatment z. In other words, it is the spillover effect from j to i, holding i's treatment fixed at z. Formally, we write

$$\xi_{ij}(z) = \mathrm{E}\big[y_{ij}(z,1;\mathbf{Z}_{-ij}) - y_{ij}(z,0;\mathbf{Z}_{-ij})\big],$$

where, on analogy with $\mathbf{Z}_{-i}$,

$$\mathbf{Z}_{-ij} = (Z_1,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_{j-1},Z_{j+1},\ldots,Z_n)$$

is the treatment vector with the ith and jth elements deleted, and $y_{ij}(a,b;\mathbf{z}_{-ij})$ is unit i's potential outcome when units i and j are assigned to treatments a and b, respectively, and the assignments of the remaining units are $\mathbf{z}_{-ij}$.

To describe the overall spillover effect between two units, consider

$$\breve{\xi}_{ij} = \mathrm{E}\big[\xi_{ij}(1-Z_i)\big] = (1-p_i)\xi_{ij}(1) + p_i\xi_{ij}(0),$$

where $p_i = \Pr(Z_i = 1)$ as above. This is the expected spillover effect using the opposite probabilities for i's treatment. That is, if i has a high probability of being assigned $Z_i = 1$, so that p_i is close to one, then $\breve{\xi}_{ij}$ gives more weight to $\xi_{ij}(0)$, which is the spillover effect when $Z_i = 0$. Let

$$\breve{Y}_i = (1-p_i)\,\mathrm{E}[Y_i \mid Z_i = 1] + p_i\,\mathrm{E}[Y_i \mid Z_i = 0]$$

denote the same type of average for the outcome of unit i. These types of quantities occasionally appear in variances of design-based estimators. The "tyranny of the minority" estimator introduced by Lin (2013) is one such example.

PROPOSITION 7. *Under a Bernoulli design and Assumption 1 with $q \geq 4$,*

$$nd_{\text{AVG}}^{-1}\big[\widehat{\mathrm{Var}}_{\text{BER}}(\hat{\tau}_{\text{HT}}) - \mathrm{Var}(\hat{\tau}_{\text{HT}})\big] \xrightarrow{p} \frac{\tau_{\text{MSQ}}}{d_{\text{AVG}}} - B_1 - B_2,$$

where

$$B_1 = \frac{1}{nd_{\text{AVG}}}\sum_{i=1}^{n}\sum_{j\neq i}\big(\breve{\xi}_{ij}\breve{\xi}_{ji} + 2\breve{Y}_j\big[\xi_{ij}(1) - \xi_{ij}(0)\big]\big),$$

$$B_2 = \frac{1}{nd_{\text{AVG}}}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{a=0}^{1}\sum_{b=0}^{1}(-1)^{a+b}\,\mathrm{Cov}(Y_i, Y_j \mid Z_i = a, Z_j = b).$$

The proposition extends the limit result in equation (1) to settings with interference. Indeed, as shown by Corollary A5 in Supplement A (Sävje, Aronow and Hudgens (2021)), the limit under no interference is a special case of Proposition 7. The relevant scaling under interference is nd_{AVG}^{-1} rather than n, which accounts for the fact that the variance may diminish at a slower rate. In particular, the scaling ensures that $nd_{\text{AVG}}^{-1}\,\mathrm{Var}(\hat{\tau}_{\text{HT}})$ is on a constant

scale. The constant q in Assumption 1 is now required to be at least four, as is typical for convergence of variance estimators.

Compared to the setting with no interference, the limit of the variance estimator contains two additional terms. The term B_1 captures the consequences of direct interference between the units. That is, it captures whether unit i interferes with unit j directly, in which case $\breve{\xi}_{ji} \neq 0$. If there is no interference, there are no spillover effects, so $\xi_{ij}(1) = \xi_{ij}(0) = 0$, and B_1 is zero. The term B_2 captures the consequences of indirect interference between units, namely when a third unit interferes with both i and j.

While any of the three terms of the limit in Proposition 7 can dominate the others asymptotically, B_2 will generally be the one we need to worry about. To see this, observe that τ_{MSQ} is negligible on a normalized scale as long as the average interference dependence is not bounded asymptotically: $d_{\mathrm{AVG}} \to \infty$. The amount of direct interference dependence is given by C_1, so $B_1 = \mathcal{O}(d_{\mathrm{AVG}}^{-1} C_1)$. The amount of indirect dependence is given by d_{AVG}, so $B_2 = \mathcal{O}(1)$. Hence, B_2 is the dominating term whenever d_{AVG} dominates C_1, which, as we noted in Section 5.2, is the case whenever the interference is not too tightly clustered.

The key insight here is that when there is interference, these two additional terms are generally nonzero, and their sum may be both positive and negative. As a consequence, the variance estimator may be asymptotically anticonservative, painting an overly optimistic picture about the precision of the point estimator. In other words, the use of conventional variance estimators under interference could be misleading.

6.2. *Alternative estimators*. The route we will explore to account for the potential anticonservativeness is to inflate the conventional estimator with various measures of the amount of interference. In addition to providing a simple way to construct a reasonable variance estimator when these measures are known, or presumed to be known, this route facilitates constructive discussions about the consequences of interference even when the measures are not known. In particular, a sensitivity analysis becomes straightforward because the conventional variance estimate is simply multiplied by the sensitivity parameter.

We will not derive the limits of these modified variance estimators. Indeed, such limits do not always exist. We will instead focus on obtaining the main property we seek, namely conservativeness in large samples. This is formalized by a one-sided consistency property, as described in the following definition.

DEFINITION 10. A variance estimator $\hat{V}$ is said to be *asymptotically conservative* with respect to the variance of $\hat{\tau}_{\mathrm{HT}}$ if

$$\lim_{\varepsilon \to 0^+} \lim_{n \to \infty} \Pr\bigl(n d_{\mathrm{AVG}}^{-1} \bigl[\hat{V} - \mathrm{Var}(\hat{\tau}_{\mathrm{HT}})\bigr] \leq -\varepsilon\bigr) = 0.$$

The interference measure that first might come to mind is the average interference dependence, d_{AVG}, which we used for the results in Section 5. Using this quantity for the inflation, we get the estimator

$$\widehat{\mathrm{Var}}_{\mathrm{AVG}}(\hat{\tau}_{\mathrm{HT}}) = d_{\mathrm{AVG}} \widehat{\mathrm{Var}}_{\mathrm{BER}}(\hat{\tau}_{\mathrm{HT}}).$$

This will generally not inflate the estimator enough to ensure conservativeness. The issue is that the interference structure could couple with the potential outcomes in such a way that the interference introduces dependence between units with large outcomes. Using d_{AVG} for the inflation requires that no such coupling takes place, or that any coupling is asymptotically negligible. The following proposition formalizes the result.

PROPOSITION 8. *The variance estimator* $\widehat{\mathrm{Var}}_{\mathrm{AVG}}(\hat{\tau}_{\mathrm{HT}})$ *is asymptotically conservative under a Bernoulli design if Assumptions 1 and 2 hold with* $q \geq 4$ *and* $\mathrm{SD}_{\sigma^2} = \mathrm{o}(d_{\mathrm{RMS}}^{-1} d_{\mathrm{AVG}})$, *where*

$$\mathrm{SD}_{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\sigma_i^2 - \frac{1}{n}\sum_{j=1}^{n}\sigma_j^2\right]^2}, \qquad \sigma_i^2 = \mathrm{Var}\left(\frac{Z_i Y_i}{p_i} - \frac{(1-Z_i)Y_i}{1-p_i}\right),$$

and

$$d_{\mathrm{RMS}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2}.$$

The condition $\mathrm{SD}_{\sigma^2} = \mathrm{o}(d_{\mathrm{RMS}}^{-1} d_{\mathrm{AVG}})$ is the design-based equivalent of a homoscedasticity assumption. In particular, σ_i^2 is the unit-level contribution to the variance of the point estimator, so SD_{σ^2} is the standard deviation of the unit-level variances. The quantity d_{RMS} is the root mean square of d_i, and d_{AVG} is the average, so $d_{\mathrm{RMS}}^{-1} d_{\mathrm{AVG}} \leq 1$. The condition thus states that SD_{σ^2} diminishes quickly, requiring that the unit-level variances are approximately the same. When this is the case, no coupling of consequence can occur, so the inflated estimator is conservative.

The homoscedasticity condition in Proposition 8 is strong, and it will generally not hold. When it does not, the estimator must be further inflated. In particular, to capture possible coupling, the inflation factor must take into account the skewness of the unit-level interference dependencies. A straightforward way to account for such skewness is to substitute the maximum for the mean, producing the estimator

$$\widehat{\mathrm{Var}}_{\mathrm{MAX}}(\hat{\tau}_{\mathrm{HT}}) = d_{\mathrm{MAX}}\,\widehat{\mathrm{Var}}_{\mathrm{BER}}(\hat{\tau}_{\mathrm{HT}}),$$

where d_{MAX} is the maximum of d_i over $i \in \mathbf{U}$.

PROPOSITION 9. *The variance estimator* $\widehat{\mathrm{Var}}_{\mathrm{MAX}}(\hat{\tau}_{\mathrm{HT}})$ *is asymptotically conservative under a Bernoulli design if either*:

A. *Assumption 1 holds with* $q \geq 4$ *and* $d_{\mathrm{MAX}} = \mathrm{o}(n^{0.5} d_{\mathrm{AVG}}^{0.5})$, *or*
B. *Assumptions 1 and 2 hold with* $q \geq 4$ *and* $\tau_{\mathrm{MSQ}} = \Omega(1)$.

The proposition demonstrates that the conventional variance estimator inflated with d_{MAX} is conservative without a homoscedasticity assumption. Part A of the proposition stipulates that d_{MAX} is dominated by the geometric mean of n and d_{AVG}, implying that the maximum of d_i does not grow too quickly compared to the average. The condition ensures that the inflated estimator concentrates. As noted above, however, an estimator can be asymptotically conservative even when it is not convergent. The concern in that case is that part of the sampling distribution may approach zero at a faster rate than the growth rate of the inflation factor, leading to anticonservativeness. Part B of the proposition provides sufficient conditions to avoid such behavior, namely that τ_{MSQ} is asymptotically bounded from below. This lower bound ensures that the unit-level treatment effects do not concentrate around zero, implying either that the average treatment effect is not zero or that there is some effect heterogeneity.

Using d_{MAX} for the inflation will generally be too conservative. At the expense of some additional complexity, we can construct a variance estimator that uses an inflation factor in-between the average and the maximum. Let $\mathbf{D}$ be a matrix whose typical argument is d_{ij}, and let $\lambda_{\max}$ be the largest eigenvalue, or spectral radius, of this matrix. One can interpret $\mathbf{D}$ as the adjacency matrix of a graph in which the units are vertices and d_{ij} denotes whether

there is an edge between i and j. The largest eigenvalue acts as a measure of the amount of interference in the sense that $\lambda_{\max} = 1$ when there is no interference, and $\lambda_{\max} = n$ when all units are interference dependent. Additionally, $\lambda_{\max}$ weakly increases with the interference. Using the spectral radius as the inflation factor, the variance estimator becomes

$$\widehat{\mathrm{Var}}_{\mathrm{SR}}(\hat{\tau}_{\mathrm{HT}}) = \lambda_{\max}\widehat{\mathrm{Var}}_{\mathrm{BER}}(\hat{\tau}_{\mathrm{HT}}).$$

The spectral radius is such that $d_{\mathrm{AVG}} \leq \lambda_{\max} \leq d_{\mathrm{MAX}}$, showing that the estimator inflated by $\lambda_{\max}$ is more conservative than when inflated by d_{AVG} but less conservative than when inflated by d_{MAX}. The inflation is sufficient for conservativeness under weaker conditions than those of Proposition 9.

PROPOSITION 10. *The variance estimator* $\widehat{\mathrm{Var}}_{\mathrm{SR}}(\hat{\tau}_{\mathrm{HT}})$ *is asymptotically conservative under a Bernoulli design if either*:

A. *Assumption* 1 *holds with* $q \geq 4$ *and* $\lambda_{\max} = \mathrm{o}(n^{0.5}d_{\mathrm{AVG}}^{0.5})$, *or*
B. *Assumptions* 1 *and* 2 *hold with* $q \geq 4$ *and* $\tau_{\mathrm{MSQ}} = \Omega(1)$.

The adjustments needed to ensure conservativeness highlight that interference may introduce considerable imprecision. However, observe that the inflation factors we use in this section are constructed to accommodate the worst case. The adjusted variance estimators will often be overly conservative. Indeed, interference can improve precision, and no inflation is then required.

Improved variance estimators are possible if we have more information about the interference structure. For example, Aronow, Crawford and Zubizarreta (2018) construct a variance estimator that requires the experimenter to know d_i for all units. This estimator will generally be less conservative than those we have introduced in this section. Sharper variance estimators are also possible if larger departures from the conventional estimator are acceptable. For example, it is sufficient to use d_{RMS} as the inflation factor if higher moments of the unit-level variances are substituted for the conventional estimator. Such an estimator will often be less conservative than the estimators above because $d_{\mathrm{AVG}} \leq d_{\mathrm{RMS}} \leq \lambda_{\max}$.

6.3. *Tail bounds*. Experimenters often use variance estimates to construct bounds on the tails of the sampling distribution of the point estimator, which in turn may be used to construct confidence intervals and hypothesis tests. A common approach is to combine a conservative variance estimator with a normal approximation of the sampling distribution, motivated by a central limit theorem. Such approximations may be reasonable when the interference is very sparse. For example, Theorem 2.7 in Chen and Shao (2004) can be applied to the HT point estimator under the Bernoulli design if $d_{\mathrm{MAX}} = \mathcal{O}(1)$, showing that the sampling distribution is approximately normal in large samples. This condition is, however, stronger than Assumption 2. The following proposition demonstrates that normal approximations will generally not be accurate under the conditions considered in this paper.

PROPOSITION 11. *Chebyshev's inequality is asymptotically sharp with respect to the sampling distribution of the* HT *estimator for every sequence of Bernoulli designs under Assumptions* 1 *and* 2. *The inequality remains sharp when Assumption* 2 *is strengthened to* $d_{\mathrm{AVG}} = \mathcal{O}(1)$ *and* $d_{\mathrm{MAX}} = \mathcal{O}(n^{0.5})$.

Tail bounds based on Chebyshev's inequality are wider than those based on a normal distribution, so the proposition implies that a normal approximation is appropriate only under stronger conditions than those used for consistency in Section 5. Indeed, $d_{\mathrm{AVG}} = \mathcal{O}(1)$ was

the strongest interference condition under consideration in that section, providing root-n consistency. Proposition 11 can be extended to other designs, including complete and paired randomization, as discussed in its proof in Supplement A (Sävje, Aronow and Hudgens (2021)).

The conclusion is that Chebyshev's inequality is an appropriate way to construct confidence intervals and conduct testing when it is not reasonable to assume that d_{MAX} is much smaller than the sample size. It may be possible to prove a central limit theorem or otherwise derive less conservative tail bounds, even when d_{MAX} is large, if one imposes other types of assumptions, for example on the magnitude of the interference.

7. Other designs and external validity. By marginalizing over the experimental design, the EATE estimand captures an average treatment effect in the experiment that actually was implemented. A consequence is that the estimand may have taken a different value if another design were used. We saw an example of this in the vaccination study in Section 3.4 where the effect of the vaccine was different depending on the vaccination rate. Experimenters know that the results from a single experiment may not extend beyond the present sample. When units interfere, concerns about external validity should also include experimental designs.

In this section, we elaborate on this concern by asking to what extent the effect for one design generalizes to other designs. Because an experiment only provides information about the potential outcomes on its support, the prospects of extrapolation are limited. The hope is that the results of an experiment may be informative of the treatment effects under designs that are close to the one that was implemented.

It will be helpful to introduce notation that allows us to differentiate between the design that actually was used and an alternative design that could have been used but was not. Let P denote the probability measure of the design that was implemented, and let Q be the probability measure of the alternative design. A subscript indicates which measure various probability operators refer to. For example, $\mathrm{E}_P[Y_i]$ is the expected outcome for unit i under the P design, and $\mathrm{E}_Q[Y_i]$ is the same under the alternative design. The question we ask here is how informative

$$\tau_{\text{P-EATE}} = \mathrm{E}_P\big[\tau_{\text{ATE}}(\mathbf{Z})\big] \quad \text{is about} \quad \tau_{\text{Q-EATE}} = \mathrm{E}_Q\big[\tau_{\text{ATE}}(\mathbf{Z})\big].$$

To answer this question, we will use measures of closeness of designs. These measures are given meaning by coupling them with some type of structural assumption on the potential outcomes. The stronger these structural assumptions are, the more we can hope to extrapolate. In line with the rest of the paper, we focus on a relatively weak assumption here, effectively limiting ourselves to local extrapolation. In particular, we assume that the treatment effects are bounded.

ASSUMPTION 7 (Bounded unit-level effects). There exists a constant k_τ such that $|\tau_i(\mathbf{z}_{-i})| \leq k_\tau$ for all $i \in \mathbf{U}$ and $\mathbf{z} \in \{0, 1\}^n$.

Bounded treatment effects are not implied by the regularity conditions in Assumption 1. These conditions ensure that the potential outcomes are well behaved with respect to the design that was actually implemented. They do not ensure that the potential outcomes are well behaved with respect to other designs. From this perspective, Assumption 7 can be seen as an extension of Assumption 1C, ensuring that the potential outcomes are well behaved for all designs.

It remains to define a measure of closeness of designs. A straightforward choice is the total variation distance between the distributions of the designs:

$$\delta(P, Q) = \sup_{x \in \mathcal{F}} \big|P(x) - Q(x)\big|,$$

where $\mathcal{F}$ is the event space of the designs, typically the power set of $\{0, 1\}^n$. The total variation distance is the largest difference in probability between the designs for any event in the event space. It is an unforgiving measure in the sense that it defines closeness purely as overlap between the two distributions. Its advantage is that it requires little structure on the potential outcomes to be informative. In particular, Assumption 7 implies that the assignment-conditional average treatment effect function $\tau_{\text{ATE}}(\mathbf{z})$ is bounded, which gives the following result.

PROPOSITION 12. *Given Assumption 7,*

$$|\tau_{\text{P-EATE}} - \tau_{\text{Q-EATE}}| \leq 2k_\tau \delta(P, Q).$$

The intuition behind the proposition is that if the two designs overlap to a large degree, then the marginalization over $\tau_{\text{ATE}}(\mathbf{z})$ will overlap to a similar degree. The proposition demonstrates that an experiment can remain informative beyond the current design under relatively weak assumptions. However, given the unforgiving nature of the total variation distance, the bound says little more than that the designs must be close to identical to be informative of each other. An example illustrates the concern.

The example compares the estimand under a Bernoulli design with $p_i = 1/2$ for all units with the estimand under complete randomization with $p = 1/2$. Consider the event $\sum_{i=1}^n Z_i = \lfloor n/2 \rfloor$. By construction of the complete randomization design, the probability of this event is one. Under Bernoulli randomization, the probability approaches zero:

$$\frac{1}{2^n}\binom{n}{\lfloor n/2 \rfloor} = \mathcal{O}(n^{-0.5}).$$

Hence, the total variation distance between the designs approaches one. When taken at face value, this means that EATE under one of the designs provides no more information about the estimand under the other design than what already is provided by Assumption 7.

We can sharpen the bound if we know that the interference is limited. There are several ways to take advantage of a sparse interference structure when extrapolating between designs. The route we explore here is to consider how sparseness affects the average treatment effect function, as captured in the following lemma.

LEMMA 2. *Given Assumption 7, $\tau_{\text{ATE}}(\mathbf{z})$ is $2k_\tau n^{-1/r} C_{r/(r-1)}$-Lipschitz continuous with respect to the L_r distance over $\{0, 1\}^n$ for any $r \geq 1$.*

The lemma says that $\tau_{\text{ATE}}(\mathbf{z})$ does not change too quickly in $\mathbf{z}$ under sparse interference. The intuition is that changing a unit's treatment can affect only a limited number of other units when the interference is sparse, and the bound on the unit-level treatment effects limits how consequential the change can be on the affected units. The lemma uses $C_{r/(r-1)}$ to measure the amount of interference rather than d_{AVG}. This is the $r/(r-1)$-norm of the unit-level interference count defined in Section 4, where these quantities were used to bound d_{AVG}. The lemma can be sharpened if more information about the potential outcomes is available. For example, the factor $2k_\tau$ can be reduced if we replace Assumption 7 with a Lipschitz continuity assumption directly on unit-level effects.

Lemma 2 is useful because the L_r distance provides more information about the similarity of different assignments than the discrete metric implicit in the total variation distance. However, to take advantage of this information, we must modify the measure of design closeness to incorporate the geometry given by the L_r distance. The Wasserstein metric accomplishes this, and it couples well with Lipschitz continuity.

Let $\mathcal{J}(P, Q)$ collect all distributions $(\mathbf{Z}', \mathbf{Z}'')$ over $\{0, 1\}^n \times \{0, 1\}^n$ such that the marginal distributions of $\mathbf{Z}'$ and $\mathbf{Z}''$ are P and Q, respectively. Each distribution in $\mathcal{J}(P, Q)$ can be seen as a way to transform P to Q by moving mass between the $\{0, 1\}^n$ points in the marginal distributions. The Wasserstein metric is the least costly way to make this transformation with respect to the L_r distance:

$$W_r(P, Q) = \inf_{J \in \mathcal{J}(P,Q)} \mathrm{E}_J\big[\|\mathbf{Z}' - \mathbf{Z}''\|_r\big],$$

where the subscript denotes the underlying L_r distance rather than the order of the Wasserstein metric, which is taken to be one here. This metric is more forgiving than the total variation distance because it goes beyond direct overlap and also considers how close the nonoverlapping parts of the distributions are. An application of the Kantorovich–Rubinstein duality theorem (Edwards (2011)) provides the following result.

PROPOSITION 13. *Given Assumption 7,*

$$|\tau_{\text{P-EATE}} - \tau_{\text{Q-EATE}}| \leq 2k_\tau n^{-1/r} C_{r/(r-1)} W_r(P, Q).$$

The proposition provides the central insight of this section. Namely, if the amount of interference does not grow too quickly relative to the difference between the designs as captured by the Wasserstein metric,

$$C_{r/(r-1)} W_r(P, Q) = \mathrm{o}\big(n^{1/r}\big),$$

then the expected average treatment effects under the two designs will converge. The optimal choice of r depends on how unevenly the interference is distributed among the units. For example, if the interference is skewed, then $r = 2$ may be reasonable to avoid being sensitive to the outliers. In that case, $C_{r/(r-1)} = C_2$ is the root mean square of the interference count. If all units interfere with approximately the same number of other units, then $r = 1$ is a better choice, in which case $C_{r/(r-1)}$ is taken to be C_∞.

Our example with the Bernoulli and complete randomization designs provides a good illustration of how Proposition 13 allows us to generalize the results from one design to another. When $r = 2$, the corresponding Wasserstein distance between the designs is $W_2(P, Q) = \mathcal{O}(n^{0.25})$. It follows that the two estimands converge as long as $C_2 = \mathrm{o}(n^{0.25})$. When $r = 1$, the corresponding Wasserstein distance is $W_1(P, Q) = \mathcal{O}(n^{0.5})$, so the estimands converge whenever $C_\infty = \mathrm{o}(n^{0.5})$.

The proposition may also prove useful when the estimands do not converge. For example, we have $n^{-1/r} W_r(P, Q) \sim |p - q|^{1/r}$ when P and Q are two complete randomization designs with p and q as their respective assignment probabilities. The corresponding estimands will generally not converge, but Proposition 13 provides a useful bound if $C_{r/(r-1)}$ and $|p - q|$ are reasonably small. The bound becomes more informative when more is known about the potential outcomes so that the factor $2k_\tau$ can be reduced.

8. Simulation study. Supplement B (Sävje, Aronow and Hudgens (2021)) presents the results from a simulation study that illustrates and complements the results presented here. We include three types of data generating processes in the simulations, differing in the structure of the interference. In particular, we investigate when the interference is contained within groups of units, when the interference structure is randomly generated, and when only one unit is interfering with other units. For each type of interference structure, we alter the amount of interference to range from $d_{\text{AVG}} = 1$ to $d_{\text{AVG}} = n$.

The simulation study corroborates the theoretical results. The estimators approach EATE at the rates given by the propositions in Section 5, and they generally do not converge when

Assumption 2 does not hold. There are, however, situations where they converge at faster rates than those guaranteed by the propositions, which highlights that the theoretical results focus on the worst case given the stated conditions. For example, in one instance, the experimental design aligns with the interference structure in such a way that the design almost perfectly counteracts the interference. The precision of the estimator does not significantly depend on the amount of interference in this case. The setting is rather artificial, however, and it was selected to illustrate exactly this point. A slight modification of the data generating process makes the estimator sensitive to the amount of interference again. We direct readers to the supplementary material for further insights from the simulation study.

9. Concluding remarks. Experimenters worry about interference. The first line of defense tends to be to design experiments in a way that minimizes the risk that units will interfere. One could, for example, physically isolate the units throughout the study. The designs needed to rule out interference may, however, make the experiments so alien to the topics under study that the findings are no longer relevant. The results would not generalize to the real world where units do interfere. When design-based fixes are undesirable or incomplete, one could try to account for any lingering interference in the analysis, but doing so requires detailed knowledge about its structure. The typical experimenter neither averts all interference by design nor accounts for it in the analysis. Instead, they conduct and analyze the experiment as if no units interfere, even when the no-interference assumption at best holds only approximately. The disconnect between assumptions and reality is reconciled by what appears to be a common intuition among experimenters that goes against the conventional view: unmodeled interference is not a fatal flaw so long as it is limited. The results in this paper provide rigorous justification for this intuition.

The EATE estimand generalizes the average treatment effect to experiments with interference, but some interpretations of ATE do not apply. In particular, EATE cannot be interpreted as the difference between the average outcome when no unit is treated and the average outcome when all units are treated. The estimand is instead the expected average effect of changing a single treatment in the current experiment. From a practical perspective, these marginal effects are relevant to policy makers considering decisions along an intensive margin. From a theoretical perspective, EATE could act as a sufficient statistic for a structural model, thereby allowing researchers to pin down theoretically important parameters (Chetty (2009)).

The main purpose of the estimand is, however, to describe what can be learned from an experiment under unknown and arbitrary interference. As shown by Basse and Airoldi (2018b) and others, causal inference under interference generally requires strong assumptions. The consistency results in the paper nevertheless show that experiments often are informative of EATE even in the presence of moderate inference with unknown form. This insight is valuable even when EATE is not the parameter of primary interest because it shows what can be learned from an experiment without imposing strong structural assumptions. A comparison can here be made with the *local average treatment effect* (LATE) estimand for the instrumental variable estimator (Imbens and Angrist (1994)). The local effect may not be the parameter of primary interest, but it is relevant because it describes what can be learned in experiments with noncompliance without strong assumptions about, for example, constant treatment effects.

We conjecture that the results in this paper extend also to observational studies. Several issues must, however, be addressed before this question can be investigated formally. These issues are mainly conceptual in nature. We do not know of a stochastic framework that can accommodate unknown and arbitrary interference in an observational setting because the design (or the assignment mechanism as it is often called in an observational setting) is then unknown. A common way to approach this problem is to approximate the assignment mechanism with a design that is easy to analyze, such as the Bernoulli design. Under an assumption

that the units' marginal treatment probabilities are given by a function depending only on the units' own characteristics, the remaining properties of the assignment mechanism may be estimated from the data. The concern with this approach is that the behavior of the estimators is sensitive to details of the design, as shown in Proposition 5, so the approximation may not be appropriate. Furthermore, the treatment probabilities may depend on the characteristics of other units, effectively capturing interference in treatment assignment. Forastiere, Airoldi and Mealli (2017) address these concerns by assuming that a unit's treatment probability is a function of both its own characteristics and the characteristics of its neighbors in an interference graph. This approach cannot be used here, however, because it requires that the interference structure is known.

Another potential way to extend the results to observational studies is to expand the stochastic framework to include sampling variability. The sample is then assumed to be randomly drawn from some larger, possibly infinite, superpopulation. When interference is investigated in this type of regime, the units are often assumed to be sampled in such a way as to maintain the interference structure. However, the only way to ensure that the interference structure is maintained under arbitrary interference is to consider the whole sample being sampled jointly, which invalidates the use of conventional proof strategies. The construction of a stochastic framework for observational studies will require close attention to these issues, but we see no reason why a marginalization argument similar to the one in this paper would not apply once an appropriate framework has been constructed.

We have focused on the effect of a unit's own treatment in this paper. The results are, however, not necessarily restricted to primary or direct treatment effects as typically defined. In particular, the pairing between units and treatments is arbitrary in our causal model, and an experiment could have several reasonable pairings. Consider the vaccination example in Section 3.4. The most natural pairing might be to let a unit's treatment indicator denote whether the unit itself was vaccinated. However, nothing prohibits us from letting it denote whether some other unit in the sample was vaccinated. For example, we could let z_i denote whether unit i's spouse was vaccinated, in which case EATE would capture the expected spillover effect between spouses. In this sense, the current investigation applies both to usual treatment effects and to rudimentary spillover effects. We conjecture that the results can be extended to other definitions of treatment, and if so, they would provide robustness to estimators of more intricate spillover effects under unknown and arbitrary interference.

Acknowledgements. We thank Alexander D'Amour, Matthew Blackwell, David Choi, Forrest Crawford, Peng Ding, Naoki Egami, Avi Feller, Owen Francis, Elizabeth Halloran, Lihua Lei, Cyrus Samii, Jasjeet Sekhon and Daniel Spielman for helpful comments and discussions. Michael Hudgens was supported by NIH Grant R01 AI085073. A previous version of this article was circulated under the title "A folk theorem on interference in experiments."

SUPPLEMENTARY MATERIAL

Supplement A: Proofs (DOI: 10.1214/20-AOS1973SUPPA; .pdf). Proofs for all propositions in the manuscript.

Supplement B: Simulation study (DOI: 10.1214/20-AOS1973SUPPB; .pdf). Results from a simulation study illustrating the results in the paper.

Supplement C: Code for simulation study (DOI: 10.1214/20-AOS1973SUPPC; .zip). R code to replicate the results in the simulation study.

REFERENCES

ANGRIST, J. D. (2014). The perils of peer effects. *Labour Econ*. **30** 98–108. https://doi.org/10.1016/j.labeco.2014.05.008

ARONOW, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociol. Methods Res*. **41** 3–16. MR3190698 https://doi.org/10.1177/0049124112437535

ARONOW, P. M., CRAWFORD, F. W. and ZUBIZARRETA, J. R. (2018). Confidence intervals for linear unbiased estimators under constrained dependence. *Electron. J. Stat*. **12** 2238–2252. MR3830833 https://doi.org/10.1214/18-EJS1448

ARONOW, P. M. and SAMII, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat*. **11** 1912–1947. MR3743283 https://doi.org/10.1214/16-AOAS1005

ATHEY, S., ECKLES, D. and IMBENS, G. W. (2018). Exact p-values for network interference. *J. Amer. Statist. Assoc*. **113** 230–240. MR3803460 https://doi.org/10.1080/01621459.2016.1241178

BASSE, G. W. and AIROLDI, E. M. (2018a). Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika* **105** 849–858. MR3877869 https://doi.org/10.1093/biomet/asy036

BASSE, G. W. and AIROLDI, E. M. (2018b). Limitations of design-based causal inference and A/B testing under arbitrary and network interference. *Sociol. Method.* **48** 136–151. https://doi.org/10.1177/0081175018782569

BASSE, G. and FELLER, A. (2018). Analyzing two-stage experiments in the presence of interference. *J. Amer. Statist. Assoc*. **113** 41–55. MR3803438 https://doi.org/10.1080/01621459.2017.1323641

BASSE, G. W., FELLER, A. and TOULIS, P. (2019). Randomization tests of causal effects under interference. *Biometrika* **106** 487–494. MR3949317 https://doi.org/10.1093/biomet/asy072

BLUM, J. R., HANSON, D. L. and KOOPMANS, L. H. (1963). On the strong law of large numbers for a class of stochastic processes. *Z. Wahrsch. Verw. Gebiete* **2** 1–11. MR0161369 https://doi.org/10.1007/BF00535293

BOWERS, J., FREDRICKSON, M. M. and PANAGOPOULOS, C. (2013). Reasoning about interference between units: A general framework. *Polit. Anal*. **21** 97–124. https://doi.org/10.1093/pan/mps038

BRAMOULLÉ, Y., DJEBBARI, H. and FORTIN, B. (2009). Identification of peer effects through social networks. *J. Econometrics* **150** 41–55. MR2525993 https://doi.org/10.1016/j.jeconom.2008.12.021

CHEN, L. H. Y. and SHAO, Q.-M. (2004). Normal approximation under local dependence. *Ann. Probab.* **32** 1985–2028. MR2073183 https://doi.org/10.1214/009117904000000450

CHETTY, R. (2009). Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. *Ann. Rev. Econ*. **1** 451–488. https://doi.org/10.1146/annurev.economics.050708.142910

CHOI, D. (2017). Estimation of monotone treatment effects in network experiments. *J. Amer. Statist. Assoc*. **112** 1147–1155. MR3735366 https://doi.org/10.1080/01621459.2016.1194845

COX, D. R. (1958). *Planning of Experiments*. Wiley, New York; CRC Press, London. MR0095561

ECK, D. J., MOROZOVA, O. and CRAWFORD, F. W. (2018). Randomization for the direct effect of an infectious disease intervention in a clustered study population. Available at arXiv:1808.05593.

ECKLES, D., KARRER, B. and UGANDER, J. (2016). Design and analysis of experiments in networks: Reducing bias from interference. *J. Causal Inference* **5**. https://doi.org/10.1515/jci-2015-0021

EDWARDS, D. A. (2011). On the Kantorovich–Rubinstein theorem. *Expo. Math*. **29** 387–398. MR2861765 https://doi.org/10.1016/j.exmath.2011.06.005

EGAMI, N. (2017). Unbiased estimation and sensitivity analysis for network-specific spillover effects: Application to an online network experiment. Available at arXiv:1708.08171.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, London.

FOGARTY, C. B. (2018). On mitigating the analytical limitations of finely stratified experiments. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **80** 1035–1056. MR3874309 https://doi.org/10.1111/rssb.12290

FORASTIERE, L., AIROLDI, E. M. and MEALLI, F. (2017). Identification and estimation of treatment and interference effects in observational studies on networks. Available at arXiv:1609.06245.

GOLDSMITH-PINKHAM, P. and IMBENS, G. W. (2013). Social networks and the identification of peer effects. *J. Bus. Econom. Statist*. **31** 253–264. MR3173674 https://doi.org/10.1080/07350015.2013.801251

GRAHAM, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica* **76** 643–660. MR2406869 https://doi.org/10.1111/j.1468-0262.2008.00850.x

GREEN, D. P. and GERBER, A. S. (2004). *Get Out the Vote*: *How to Increase Voter Turnout*. Brookings Institution Press, Washington, DC.

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242 https://doi.org/10.2307/2998560

HÁJEK, J. (1971). Comment: An essay on the logical foundations of survey sampling, part one. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) Holt, Rinehart and Winston, Toronto.

HALLORAN, M. E. and HUDGENS, M. G. (2016). Dependent happenings: A recent methodological review. *Curr. Epidemiol. Rep*. **3** 297–305. https://doi.org/10.1007/s40471-016-0086-4

HALLORAN, M. E. and STRUCHINER, C. J. (1995). Causal inference in infectious diseases. *Epidemiology* **6** 142–151. https://doi.org/10.2307/3702315

HERNÁN, M. A. and ROBINS, J. M. (2006). Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* **60** 578–586. https://doi.org/10.1136/jech.2004.029496

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826 https://doi.org/10.1111/1468-0262.00442

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475. https://doi.org/10.2307/2951620

ISAKI, C. T. and FULLER, W. A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77** 89–96. MR0648029

JAGADEESAN, R., PILLAI, N. and VOLFOVSKY, A. (2017). Designs for estimating the treatment effect in networks with interference. Available at arXiv:1705.08524.

KANG, H. and IMBENS, G. W. (2016). Peer encouragement designs in causal inference with partial interference and identification of local average network effects. Available at arXiv:1609.04464.

LEE, L. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *J. Econometrics* **140** 333–374. MR2408910 https://doi.org/10.1016/j.jeconom.2006.07.001

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420 https://doi.org/10.1214/12-AOAS583

LIU, L. and HUDGENS, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *J. Amer. Statist. Assoc.* **109** 288–301. MR3180564 https://doi.org/10.1080/01621459.2013.844698

LIU, L., HUDGENS, M. G. and BECKER-DREPS, S. (2016). On inverse probability-weighted estimators in the presence of interference. *Biometrika* **103** 829–842. MR3620442 https://doi.org/10.1093/biomet/asw047

LUO, X., SMALL, D. S., LI, C.-S. R. and ROSENBAUM, P. R. (2012). Inference with interference between units in an fMRI experiment of motor inhibition. *J. Amer. Statist. Assoc.* **107** 530–541. MR2980065 https://doi.org/10.1080/01621459.2012.655954

MANSKI, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Rev. Econ. Stud.* **60** 531–542. MR1236836 https://doi.org/10.2307/2298123

MANSKI, C. F. (2013). Identification of treatment response with social interactions. *Econom. J.* **16** S1–S23. MR3030060 https://doi.org/10.1111/j.1368-423X.2012.00368.x

NICKERSON, D. W. (2008). Is voting contagious? Evidence from two field experiments. *Am. Polit. Sci. Rev.* **102** 49–57. https://doi.org/10.1017/S0003055408080039

OGBURN, E. L. and VANDERWEELE, T. J. (2017). Vaccines, contagion, and social networks. *Ann. Appl. Stat.* **11** 919–948. MR3693552 https://doi.org/10.1214/17-AOAS1023

RIGDON, J. and HUDGENS, M. G. (2015). Exact confidence intervals in the presence of interference. *Statist. Probab. Lett.* **105** 130–135. MR3371989 https://doi.org/10.1016/j.spl.2015.06.011

ROBINSON, P. M. (1982). On the convergence of the Horvitz–Thompson estimator. *Aust. J. Stat.* **24** 234–238. MR0678263 https://doi.org/10.1111/j.1467-842x.1982.tb00829.x

ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. MR2345537 https://doi.org/10.1198/016214506000001112

ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA* **42** 43–47. MR0074711 https://doi.org/10.1073/pnas.42.1.43

RUBIN, D. B. (1980). Comment: Randomization analysis of experimental data: The Fisher randomization test. *J. Amer. Statist. Assoc.* **75** 591. https://doi.org/10.2307/2287653

SÄVJE, F., ARONOW, P. M and HUDGENS, M. G (2021). Supplements to "Average treatment effects in the presence of unknown interference." https://doi.org/10.1214/20-AOS1973SUPPA, https://doi.org/10.1214/20-AOS1973SUPPB, https://doi.org/10.1214/20-AOS1973SUPPC

SINCLAIR, B., MCCONNELL, M. and GREEN, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *Amer. J. Polit. Sci.* **56** 1055–1069. https://doi.org/10.1111/j.1540-5907.2012.00592.x

SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc.* **101** 1398–1407. MR2307573 https://doi.org/10.1198/016214506000000636

SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. MR1092986

SUSSMAN, D. L. and AIROLDI, E. M. (2017). Elements of estimation theory for causal effects in the presence of network interference. Available at arXiv:1702.03578.

TCHETGEN TCHETGEN, E. J., FULCHER, I. and SHPITSER, I. (2019). Auto-g-computation of causal effects on a network. Available at arXiv:1709.01577.

TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. MR2867538 https://doi.org/10.1177/0962280210386779

TOULIS, P. and KAO, E. (2013). Estimation of causal peer influence effects. In *Proceedings of the* 30*th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.). *Proceedings of Machine Learning Research* **28** 1489–1497. PMLR, Atlanta, GA.

UGANDER, J., KARRER, B., BACKSTROM, L. and KLEINBERG, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD* '13 329–337. ACM, New York. https://doi.org/10.1145/2487575.2487695

VANDERWEELE, T. J. and TCHETGEN TCHETGEN, E. J. (2011). Effect partitioning under interference in two-stage randomized vaccine trials. *Statist. Probab. Lett.* **81** 861–869. MR2793754 https://doi.org/10.1016/j.spl.2011.02.019