Political Analysis (2026), **0**, 1–16
doi:10.1017/pan.2026.10045

PA

ARTICLE

On the Foundations of the Design-Based Approach

P. M. Aronow[1], Austin Jang[1] and Molly Offer-Westort[2]

[1]Statistics & Data Science and Political Science, Yale University, USA; [2]Political Science, University of Chicago Division of the Social Sciences, USA

Corresponding author: Molly Offer-Westort; Email: mollyow@gmail.com

(Received 10 June 2025; revised 3 April 2026; accepted 3 April 2026)

Abstract

The design-based paradigm may be adopted in causal inference and survey sampling when we assume Rubin's stable unit treatment value assumption (SUTVA) or impose similar frameworks. While often taken for granted, such assumptions entail strong claims about the data-generating process. We develop an alternative design-based approach: we first invoke a generalized, non-parametric model that allows for unrestricted forms of interference, such as spillover. We define an associated set of inferential targets and discuss their interpretation under SUTVA and a weaker assumption that we call the "no unmodeled revealable variation assumption" (NURVA). We then reconstruct the standard paradigm, reconsidering SUTVA at the end rather than assuming it at the beginning. Despite its similarity to SUTVA, we demonstrate the practical limitations of NURVA alone for identifying substantively interesting quantities. In so doing, we provide clarity on the nature and importance of SUTVA for applied research.

Keywords: design-based inference; causal inference; interference; exposure mappings; SUTVA

Edited by: Erin Hartman

1. Introduction

In classical survey sampling, design-based and model-based inference differ by the assumed source of stochasticity in the data-generating process.[1] In model-based inference, randomness comes from largely allegorical notions, such as sampling from an infinite super-population or a stochastic error term in the outcome model. In design-based inference, randomness lies in which units are treated or sampled under the *design* that assigns probabilities to "various subsets of the finite population" (Särndal 1978).

Design-based inference for causal queries originated with the influential work of Neyman (1990 [1923]) and, shortly after, was developed for survey sampling in Neyman (1934). Neyman-type design-based inference provides asymptotically justified estimation under a set of consistency assumptions and mild regularity conditions.[2] The framework established by Neyman was extended to include general causal inference problems including from observational data in Rubin (1974).[3] One of Rubin's contributions to the model was to coin the stable unit treatment value assumption (SUTVA). Today, SUTVA is characterized by two conditions. The first is *no interference*, as Rubin described, "If unit *i*

[1]Here, we are interested in the classical, statistical meaning of design-based inference, although the term often takes on a different meaning in the social sciences. For a reference on the social science interpretation of design-based inference, see Card (2022), Dunning (2010) and Kocher and Monteiro (2016).

[2]For a more comprehensive discussion of the history of design-based survey sampling and causal inference, refer to Section 2 of Aronow and Middleton, 2013.

[3]Fisher–Rosenbaum-type estimation is a distinct but related literature to the approach elaborated in this article that also assumes treatment assignment is the only source of randomness. For examples in the social sciences, see Sinclair, McConnell, and Green (2012) and Kalla and Broockman (2016).

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Society for Political Methodology.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

is exposed to treatment j, the observed value of Y will be Y_{ij}; that is, there is no interference between units (Cox 1958, p. 19) leading to different outcomes depending on the treatments other units received." The second is *no hidden treatment variations*; Rubin continues, "there are no versions of treatments leading to 'technical errors' (Neyman et al. 1935)." These assumptions let us leverage observations across treatment conditions to make causal claims about alternative interventions. In experimental and quasi-experimental settings, SUTVA implies that there are $N \times K$ potential outcomes, where N is the number of observations and K is the number of treatment levels, and each observation can take on K potential outcomes. In the survey sampling context, we invoke similar assumptions. Suppose that we survey n of N units in a finite population but we wish to make claims about what would have been observed had we surveyed the entire population. To do so, we often implicitly rely on a SUTVA-like assumption that there exist only N observable potential outcomes, one associated with each unit of the population.[4] This implies that, for every observation, we assume that who else, and how, we survey does not affect observable potential outcomes. While often taken for granted, this is a strong assumption and will have implications for the generalizability of causal quantities as well.

There have been a number of advances in recent years for the design-based approach to statistical inference, including for instrumental variables estimation (Borusyak, Hull, and Jaravel 2025), regression adjustment (Middleton 2018) and interference both spatially (Leung 2020) and temporally (Wang and Jetsupphasuk 2021). Additional recent design-based work develops estimation under relaxations of SUTVA in specific settings (Gao and Ding 2025; Heng, Zhang, and Feng 2025; Lu *et al.* 2025; Zhao and Ding 2022). More general methods for randomization tests under interference (Basse, Feller, and Toulis 2019) and Riesz estimators (Harshaw, Sävje, and Wang 2022) have also been proposed.

We contribute to this literature by proposing a new design-based framework for analyzing randomized trials and survey sampling. Departing from the usual reliance on Rubin's SUTVA, our approach begins with a generalized, non-parametric model that permits arbitrary forms of interference. Because unit-level potential outcomes are not generally well-defined under interference or hidden versions of treatment, we introduce design-conditional *expected potential outcomes* (EPOs) and their averages and design-conditional contrasts as an alternative (Equations (5)–(8)). Our defined contrasts coincide with the usual average treatment effects when SUTVA holds, but these targets remain well-defined as combinations of potential outcomes, even under interference or misspecification of exposures—as long as the *design* is known. We introduce a weaker, design-conditional variant of SUTVA that we call the "no unmodeled revealable variation assumption" (NURVA) (Condition 5.1) that requires stability only over the support of the design. When NURVA holds but SUTVA does not, contrasts remain causal under the design; but only SUTVA permits off-support interpretations.

This framework enables researchers to conduct more faithful analyses of complex data-generating processes that reflect real-world interactions. While off-support generalizations may interest policymakers and applied researchers, our reframing separates design-identified quantities from those targets.

In Section 2, we provide two stylized examples illustrating our framework. In Section 3, we establish the key elements of the analysis—design, population and exposure mapping. In Section 4, we define a set of new inferential targets for the design-based setting. In Section 5, we discuss the interpretation of these targets under Rubin's SUTVA as well as under NURVA. In Section 6, we discuss implications of SUTVA and NURVA for external validity. In Section 7, we discuss implications for estimation. In Section 8, we conclude.

2. Stylized examples

To motivate our discussion, we use two stylized versions of common research designs (a survey and an experiment in a setting where there is potential peer influence) to demonstrate the flexibility and generality of the proposed framework. Additional toy models that demonstrate important, but more particular, pathologies will be discussed later.

[4]See the fixed outcome model in classical survey sampling associated with Särndal (1978) or Kish (1965).

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

Example 1: Rebel group survey: A researcher has a list of 10 historical rebel groups in Colombia and asks: what is the average number of reported members today? To answer this question, they randomly order the 10 groups, survey the first three and record reported former members; unsurveyed groups are coded as having no members. Unsurveyed rebel groups can be coded arbitrarily (e.g., 0 or −99); this is inconsequential for defining the target or for estimation.

Example 2: School experiment: A researcher is interested in the effect of a pro-volunteering program on volunteering among a population of high school students. In the experimental design, they assign one-fifth of the students to the pro-volunteering arm under uniform random assignment. For each student, the researcher then measures whether they volunteered.

3. Setup

Following the frameworks of Neyman (1990 [1923]) and Särndal (1978), we formalize the design in terms of a probability triple (Ω, S, P), where the sample space, Ω, consists of all conceivable experimental (or sampling) interventions and each realized intervention, ω, refers to the vector of treatment (or sampling) assignments to all N observations. The event space is the power set of the sample space, $S = \mathcal{P}(\Omega)$ and we assume that our probability measure satisfies $\Pr[A] = \sum_{\omega \in A} p(\omega), \forall A \in S$, where $p : \Omega \to [0,1]$, with $p(\omega)$ characterizing the probability of the assignment of an intervention. We then embed potential outcomes per Rubin (1974) and adopt exposure mappings for interference per Aronow and Samii (2017) (see also Hudgens and Halloran 2008).

3.1. Design

We define the design space $\mathcal{Z} = \mathbf{Z}(\Omega)$, where $\mathbf{Z}$ is a bijective random vector $\mathbf{Z} : \Omega \to \mathbb{R}^N$. The *design* is defined as the joint p.m.f. of $\mathbf{Z}$, $f : \mathbb{R}^N \to [0,1]$. We may define all objects with respect to the design vector $\mathbf{Z}$, removing all notational dependence on ω. We assume that the design is determined by the researcher, grounding our discussion away from concerns about the uncertainty in the assignment procedure. More generally, a valid design exists in settings where the researcher is simply aware of the probabilities implied by an assignment process and the sample space, allowing for design-based inference in observational settings. This setup and notation are, so far, fairly standard in the design-based setting. However, we will use this setup to define novel estimands in Section 4. The conceptual differences between the design space and the design will be central to the interpretation of these estimands as discussed in Section 5.

For illustration, we define the design with respect to our stylized examples. In the rebel group survey, the design is the sampling procedure that the researcher uses to determine the probability of surveying each of the historical rebel groups, which consists of randomly sorting a list of the 10 groups and surveying the first three in order. One way to formally express this procedure is as follows:

$$f(\mathbf{z}) = \begin{cases} \frac{(10-3)!}{10!} & : \mathbf{z} \in \text{Sym}\left((1,2,3,0,0,0,0,0,0,0)\right) \\ 0 & : \text{otherwise,} \end{cases}$$

where $\text{Sym}(\cdot)$ represents the set/orbit of distinct permutations of a vector. The corresponding design space is $\mathcal{Z} \supseteq \{(0,0,0,0,0,0,0,0,0,0)\} \cup \text{Sym}\left((1,0,0,0,0,0,0,0,0,0)\right) \cup \text{Sym}\left((1,2,0,0,0,0,0,0,0,0)\right) \cup \cdots \cup \text{Sym}\left((1,2,3,4,5,6,7,8,9,10)\right)$.[5]

In our school experiment, the design is the procedure we use to determine the probability that each student receives the treatment arm. Letting $\lceil 0.2N \rceil$ be the number of observations in treatment, the design is

[5] In our notational shorthand for $\mathcal{Z}$, the number represents the ordinal position in which the indexed group was surveyed with 0 indicating that the group was not surveyed. For instance, $\mathbf{z} = (1,2,3,0,0,0,0,0,0,0)$ translates to surveying group one first, group two second and group three third, with no other group being surveyed. This is a different intervention than $\mathbf{z} = (3,2,1,0,0,0,0,0,0,0)$ in which we survey the same three groups, but we surveyed group three first, group two second and group one third.

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

$$f(\mathbf{z}) = \begin{cases} \frac{(\lceil 0.2N \rceil)!(N-\lceil 0.2N \rceil)!}{N!} & : \sum_{k=1}^{N} z_k = \lceil 0.2N \rceil \\ 0 & : \text{otherwise,} \end{cases}$$

where the corresponding design space is $\mathcal{Z} \supseteq \{0,1\}^N$.

3.2. Population

We follow Särndal (1978) in defining the population as the collection of units about which we wish to conduct inference. This finite population is indexed by $i = 1, \ldots, N$, where indexing is fixed. Our estimands are defined over this population on specified outcomes of interest. Unless otherwise specified, generalizations are to estimands defined over this same population. We represent the outcomes of interest associated with these units with a real-valued random vector, $\mathbf{Y}$. The random vector $\mathbf{Y}$ characterizes the *raw potential outcomes*, which can be represented for each individual i as

$$y_i(\mathbf{z}) = Y_i(\mathbf{Z}^{-1}(\mathbf{z})), \forall i \in 1, \ldots, N, \mathbf{z} \in \mathcal{Z}. \tag{1}$$

The raw potential outcome $y_i(\mathbf{z})$ for individual i is what we observe if we perform intervention $\mathbf{z}$ and measure unit i's outcome. There are a total of $N \times |\mathcal{Z}|$ raw potential outcomes in the causal setting. There are $\sum_{\mathbf{z} \in \mathcal{Z}} \sum_{z_i \in \mathbf{z}} \mathbb{1}\{z_i \neq 0\}$ substantively defined raw potential outcomes in the survey sampling setting.[6]

In our rebel group survey, the population is the 10 historical rebel groups and the outcome is the measured number of reported members. In our school experiment, the population is the N students, and the outcome is an indicator of whether a student volunteered.

3.3. Exposure mapping

While the raw potential outcome characterizes the outcome at the most granular level, we are often interested in further characterizing a unit of observation by the "exposure" that our intervention has placed them in. Examples of exposures in the experimental literature include social influence (Centola 2010), spatial exposure (Miguel and Kremer 2004) or simple direct exposure—what is most commonly considered in experimental settings under SUTVA. The *exposure mapping* relates interventions to the exposures ("treatments") experienced by the units of observation. The exposure mapping was originally defined, named and applied to the design-based setting in Aronow and Samii (2017). When exposure mappings are properly specified, that is, for each unit, a given exposure is associated with a single, stable value of the outcome, and if this is satisfied for all units, then exposures align with what Manski (2013) denotes as "effective treatments."

Formally, our exposure mapping for individual i is a function $g_i : \mathcal{Z} \to \mathbb{R}$. Since g_i is a function, each realization of the intervention associates each individual with only one exposure, although this can be easily relaxed to allow for settings where an observation is in multiple exposures, such as in multi-factorial designs. Given that we have knowledge of the support of the design vector, $\text{Supp}(\mathbf{Z})$, we can define a function $g_i|_{\text{Supp}(\mathbf{Z})} : \text{Supp}(\mathbf{Z}) \to \mathbb{R}$. We denote the N-length random exposure vector

$$\mathbf{D} = (g_1(\mathbf{Z}), g_2(\mathbf{Z}), \ldots, g_N(\mathbf{Z})). \tag{2}$$

We condition on the support of the design in $g_i|_{\text{Supp}(\mathbf{Z})}$ because while there may be some interventions in the design space, $\mathcal{Z}$, in which the exposure is not well-defined, for estimation, these are only consequential if they occur with positive probability.

In our rebel group survey, we might code a group as exposed if they had been surveyed and not exposed if they were not. This exposure mapping can be expressed mathematically as

$$g_i(\mathbf{z}) = \begin{cases} 1 & : z_i > 0 \\ 0 & : \text{otherwise.} \end{cases}, \forall i \in \{1, \ldots, 10\}, \mathbf{z} \in \text{Sym}\left((1,2,3,0,0,0,0,0,0,0)\right).$$

[6]The term *raw* refers to the fact that we are characterizing the well-defined potential outcome of unit i under the *intervention*, $\mathbf{z}$, as compared to the distribution of individual potential outcomes that may be associated with unit i under a given researcher defined individual level exposure.

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

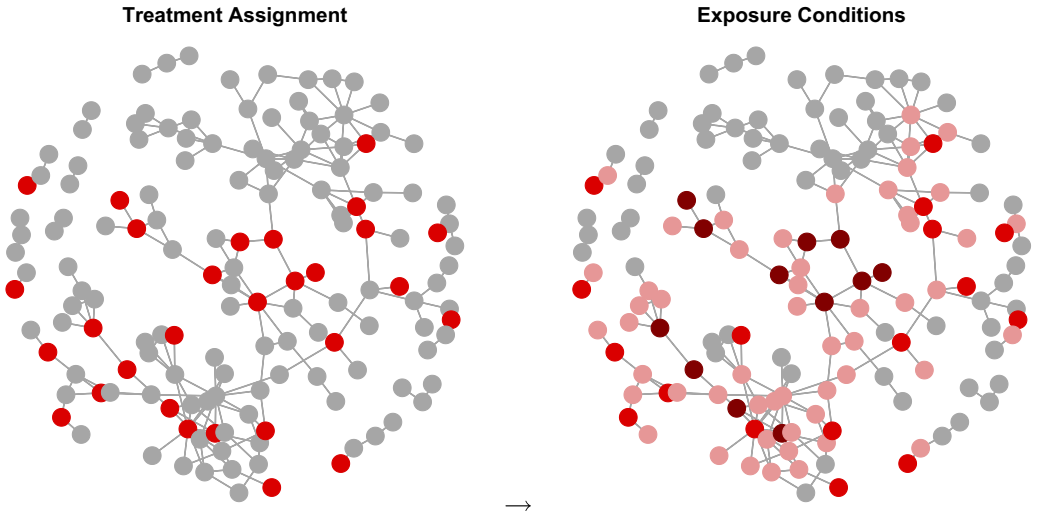


Figure 1. Network exposure. The left panel shows the intervention (treated in red, control in gray); the right panel shows resulting exposures under Equation (3): control (gray), isolated direct (red), indirect (pink) and direct plus indirect (maroon).

In our school experiment, we could use an "individualistic" exposure mapping, in which we code an individual as exposed if they were assigned to the program, and not exposed otherwise,

$$g_i(\mathbf{z}) = \begin{cases} 1 & : z_i = 1 \\ 0 & : \text{ otherwise} \end{cases}, \forall i \in \{1,\ldots,N\}, \mathbf{z} \in \{0,1\}^N.$$

Alternatively, suppose the researcher also has some pre-treatment measure of a network among peers, which may deviate from the true network in its representation of the underlying data-generating process. Using this pre-treatment measure of the social network, we could define exposures based on whether adjacent nodes were treated (Aronow and Samii 2017; Sävje 2024). One example is

$$g_i(\mathbf{z}) = \begin{cases} 0 \text{ (Control)} & : \text{if } z_i = 0 \text{ and } \theta_i\mathbf{z} = 0, \\ 1 \text{ (Isolated Direct)} & : \text{if } z_i = 1 \text{ and } \theta_i\mathbf{z} = 0, \\ 2 \text{ (Indirect)} & : \text{if } z_i = 0 \text{ and } \theta_i\mathbf{z} > 0, \\ 3 \text{ (Direct \& Indirect)} & : \text{if } z_i = 1 \text{ and } \theta_i\mathbf{z} > 0, \end{cases} \quad \forall i \in \{1,\ldots,N\},\ \mathbf{z} \in \{0,1\}^N. \tag{3}$$

where θ_i is the ith row of the network adjacency matrix.

Figure 1 illustrates this mapping: the left panel shows the intervention (treated in red, control in gray); the right panel shows resulting exposures under Equation (3): control (gray), isolated direct (red), indirect (pink) and direct plus indirect (maroon).

In Section 5, we present a toy example where the exposure mapping does not match the true underlying decision rule and discuss implications. Sävje (2024) further delineates the use of the exposure mapping to define causal estimands and for making assumptions about our potential outcomes.

4. Inferential targets

We define new inferential targets for the design-based framework. In each experimental or survey sampling setting, we observe only one realization of $(\mathbf{Z}, \mathbf{Y})$. For experimental estimands, we first define contrasts between the individual-level outcome in expectation under a design when a unit has been exposed to one condition and the outcome in expectation when that same unit has been exposed to another. We use these individual-level contrasts as building blocks for our eventual inferential targets,

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

averages of these contrasts over units in the population. Similarly, in the survey sampling setting, we define the individual-level outcome in expectation under a design when a unit has been sampled, and then construct estimands as averages across units.

Note that under this setup, the "potential outcome when unit i is in exposure d" is not generally well-defined. There may be multiple assignment vectors $\mathbf{z}$ compatible with each exposure d, and thus multiple possible raw potential outcomes. Most other work assumes this away (most commonly with Rubin's SUTVA, further discussed in Section 5), which allows for an unambiguous characterization of the mean outcome or average treatment effect. Instead, we define alternative targets that coincide with standard targets when SUTVA holds, but remain well-defined when it does not. In cases when SUTVA fails, absent alternative stability assumptions, these targets are akin to "pseudo-true" parameters, values that minimize a population-level objective function under model misspecification (Sawa 1978).

4.1. Expected potential outcome

The first target we consider is the EPO. First, we denote the marginal probability that unit i enters exposure d under the design distribution of $\mathbf{Z}$ as

$$\pi_i(d) = \left(\sum_{\mathbf{z} \in \text{Supp}(\mathbf{Z}): g_i(\mathbf{z}) = d} \Pr[\mathbf{Z} = \mathbf{z}] \right) = \Pr[D_i = d]. \tag{4}$$

We impose a condition that all observations have positive probability under the design of falling into the relevant exposure condition.

Condition 4.1 (**Individual positivity**). An exposure d satisfies *individual positivity* if there exists a c such that $0 < c \leq \pi_i(d)\ \forall i \in \{1, \ldots, N\}$.

Note that unlike in conventional approaches to inverse probability weighting, in which positivity is a population-level assumption used for estimation, here it is an individual-level assumption also used in defining the estimands. More specifically, in conventional inverse probability weighting, an individual may receive a treatment deterministically if, when drawing from the population, a given covariate profile has a non-zero probability of being observed under any treatment condition (Cole and Hernán 2008; Hernán and Robins 2006). In our setting, each individual that we conduct inference over must have a non-zero probability of being in the relevant exposure condition. While this may seem like a stringent assumption, in practice, researchers can always trim the population to a subset where the assumption is valid to arrive at interpretable functions of EPOs—although this changes the target population and therefore the estimand.

Given that relevant individual positivity assumptions hold, the EPO for unit i under exposure d is defined as

$$\begin{aligned} \bar{y}_i(d) &= \frac{\sum_{\mathbf{z} \in \text{Supp}(\mathbf{Z}): g_i(\mathbf{z}) = d} y_i(\mathbf{z}) \Pr[\mathbf{Z} = \mathbf{z}]}{\sum_{\mathbf{z} \in \text{Supp}(\mathbf{Z}): g_i(\mathbf{z}) = d} \Pr[\mathbf{Z} = \mathbf{z}]} \\ &= \mathrm{E}[y_i(\mathbf{Z}) | g_i(\mathbf{Z}) = d] \\ &= \mathrm{E}[y_i(\mathbf{Z}) | D_i = d]. \end{aligned} \tag{5}$$

This can be interpreted as the individual outcome that we would see, in expectation, when we intervene using design vector $\mathbf{Z}$ to expose unit i to exposure d.[7] To avoid notational overload, these definitions implicitly hold fixed the exposure mapping g and design $\mathbf{Z}$. They could be augmented (e.g., $\bar{y}_i(d; (g, \mathbf{Z}))$) to facilitate comparisons across different designs and exposure mappings.

[7] This is a generalization to arbitrary assignment schemes of the "individual average potential outcome" defined in Hudgens and Halloran (2008).

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

Returning to our stylized examples, in our rebel group survey with an individualistic exposure mapping, the quantity $\bar{y}_3(d=1)$ represents the number of reported historical members we would find *on average* when we survey the third rebel group using our design. While we do not typically think of surveys as requiring SUTVA-like assumptions, survey outcomes are not generally well-defined in their absence. Consider a quantity such as "the potential outcome of rebel group three, when it is surveyed." Our previously introduced exposure mapping for the survey setting implies that the number of members reported for this group should be unaffected by the act of surveying other groups (no interference) and whether the group is surveyed first, second or third (no hidden treatment variations). In our school experiment, the quantity $\bar{y}_3(d=1)$ represents the rate at which an individual student would volunteer when they are directly exposed in isolation, in expectation, under our design. In practice, we are rarely, if ever, interested in the EPO of any particular unit, but rather, we will use EPOs as building blocks to more interesting estimands.

4.2. Average expected potential outcome (AEPO)

Under individual positivity (Condition 4.1) for exposure d and all individuals $i \in \{1,\ldots,N\}$, we average the EPOs over the units in the population to define the AEPO as

$$\bar{y}(d) = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i(d) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[y_i(\mathbf{Z})|D_i = d\right]. \tag{6}$$

This is akin to asking: averaging over all units, what is the outcome that we would see, in expectation, for each respective unit when we intervene using design $\mathbf{Z}$ to expose that unit to d? In the survey setting, the AEPO of the surveyed is of particular interest, because it amounts to asking: averaging over all units, what is the outcome that we would see, in expectation, for each respective unit when they were sampled under the design $\mathbf{Z}$?

4.3. Expected exposure difference (EED)

Similarly, causal "contrasts" can be defined. For an individual observation i, assuming that Condition 4.1 holds for exposures d and d' for all individuals $i \in \{1,\ldots,N\}$, we can define the EED between d and d' as

$$\tau_i(d,d') = \bar{y}_i(d) - \bar{y}_i(d'). \tag{7}$$

This can be interpreted as the difference in individual outcomes that we would see, in expectation, when we intervene using the design vector $\mathbf{Z}$ to expose unit i to exposure d as compared to d'.[8]

4.4. Average expected exposure difference (AEED)

Just as we averaged the EPO to get the AEPO, we can average the EED to get the AEED. The AEED between d and d' is

$$\tau(d,d') = \bar{y}(d) - \bar{y}(d') = \frac{1}{N}\sum_{i=1}^{N}\tau_i(d,d'). \tag{8}$$

We can also define a causal contrast in our survey setting where the AEED is between surveyed and unsurveyed groups. However, if we do not survey a group, we do not find any former members, so the causal contrast, while defined, is not empirically tractable.

[8] We compare this to the *assignment-conditional unit-level treatment effect* defined in Sävje, Aronow, and Hudgens (2021):

$$\tau_i^{AC}(d,d',\mathbf{z}_{-i}) = y_i(d;\mathbf{z}_{-i}) - y_i(d';\mathbf{z}_{-i}).$$

In the assignment-conditional unit-level treatment effect, we hold constant treatments assigned to all other individuals at $\mathbf{z}_{-i}$. We allow the support over which we take expectations in $\mathbf{z} \in \mathrm{Supp}\left(\mathbf{Z}\right) : g_i(\mathbf{z}) = d$ and $\mathbf{z}' \in \mathrm{Supp}\left(\mathbf{Z}\right) : g_i(\mathbf{z}') = d'$ to be different. Assignment-conditional unit-level treatment effects may have a more direct individual causal interpretation than EEDs as they refer to contrasts where all else is held equal besides the individual exposure. However, the regularity conditions necessary for estimation and inference for the estimands in Sävje *et al.* (2021) can differ considerably from those invoked here.

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

5. Interpretation

Absent further stability assumptions, [A]EEDs are design-conditional contrasts: they quantify how outcomes would differ, on average, if the design induced exposure d rather than d'. These "pseudo-true" quantities that describe differences in empirical distributions of outcomes may be of interest for evaluating counterfactual real-world assignment mechanisms, even if causal interpretations do not hold at the *exposure* level. We now consider the interpretation of AEEDs under two relevant assumptions and describe when we can interpret AEEDs to be the effects of exposures themselves, and when we can interpret them in a more limited way as the effects of exposures under a particular design.

5.1. Possible assumptions

First, we consider a new assumption that we call NURVA.

Condition 5.1 (NURVA).

$$g_i(\mathbf{z}) = g_i(\mathbf{z}') \implies y_i(\mathbf{z}) = y_i(\mathbf{z}') = \overline{y}_i(g_i(\mathbf{z})), \forall \mathbf{z}, \mathbf{z}' \in \text{Supp}(\mathbf{Z}).$$

In contrast to the setting where we make no additional assumptions, under NURVA, each EPO is no longer composed of weighted averages of different possible raw potential outcomes, but instead consists of a unique and common raw potential outcome realized under any assignment $\mathbf{z} \in \text{Supp}(\mathbf{Z})$ that results in the individual being exposed to d. There is no unmodeled interference that could possibly be seen, even if we were able to observe the raw potential outcomes under every random assignment associated with the design. Under NURVA, the EED can be thought of as, "If I were to use my design to induce exposure d instead of exposure d' on that unit, how much different would the result be?" We can drop the reliance on expectations over the design.

Second, we consider the more familiar assumption of Rubin's SUTVA, which makes an assumption like NURVA, but does so across all *feasible* interventions.

Condition 5.2 (SUTVA).

$$g_i(\mathbf{z}) = g_i(\mathbf{z}') \implies y_i(\mathbf{z}) = y_i(\mathbf{z}') = \overline{y}_i(g_i(\mathbf{z})), \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}.$$

The distinction between SUTVA and NURVA hinges on how the design space of all feasible interventions is defined. In policy analysis, the inferential scope that a researcher is interested in may cover treatment or sampling interventions that go beyond those that can be assigned within a given experiment or sampling procedure, and so $\mathcal{Z}$ may include assignments outside the design. Under SUTVA, the EED can be interpreted as: "If I induce exposure d instead of exposure d' on that unit, how much different would the result be?" We can now also drop any qualifiers related to the design. This interpretation now fully coincides with that of the (average) treatment effect as it is conventionally used. For a given design, NURVA and SUTVA are observationally equivalent: NURVA embeds all observable implications of SUTVA, as all $\mathbf{z} \in \mathcal{Z} \setminus \text{Supp}(\mathbf{Z})$ remain hidden under the experimental design. Hence, the gap is in scope.

5.2. Toy example: NURVA holds but SUTVA does not

Consider a household with two persons. We flip a fair coin: heads treats only Person one; tails treats only Person two. We observe whether each votes. The exposure mapping is individualistic, $g_i(\mathbf{z}) = z_i$. The true (unknown) rule is that Person one votes $\iff$ Person two is treated, and Person two votes $\iff$ Person one is treated.

Table 1a shows revealable information. Under this design, the AEPO under treatment is $\overline{y}(1) = 0$ and under control is $\overline{y}(0) = 1$, so $\tau(1,0) = -1$. Given the design and unknown rule, the untreated person

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

Table 1. NURVA holds but SUTVA does not

(a) Revealable information				
z	**d**	$y_1(\mathbf{z})$	$y_2(\mathbf{z})$	$Pr[\mathbf{Z} = \mathbf{z}]$
(0,1)	(0,1)	1	0	0.5
(1,0)	(1,0)	0	1	0.5

$\longrightarrow \tau(1,0) = -1$

(b) Feasible information				
z	**d**	$y_1(\mathbf{z})$	$y_2(\mathbf{z})$	$Pr[\mathbf{Z} = \mathbf{z}]$
(0,0)	(0,0)	0	0	0
(0,1)	(0,1)	1	0	0.5
(1,0)	(1,0)	0	1	0.5
(1,1)	(1,1)	1	1	0

votes and the treated does not. Table 1b lists all raw potential outcomes in $\mathcal{Z} = \{0,1\}^2$, including two interventions not revealable under the design.

Here, SUTVA fails but NURVA holds. Consider interventions $(1,0)$ and $(1,1)$, where the latter is off-support in Table 1. Under the individualistic exposure mapping for Person one, these exposures are equivalent, $g_1(\mathbf{z} = (1,0)) = g_1(\mathbf{z} = (1,1)) = 1$. However, the outcomes differ, $0 = y_1(\mathbf{z} = (1,0)) \neq y_1(\mathbf{z} = (1,1)) = 1$. Therefore, SUTVA does not hold. NURVA, unaffected by off-support instability, trivially holds.

We stress that NURVA is an assumption that depends on the design, unlike SUTVA, which is concerned with the entirety of the design space. Using the same design space, we illustrate this with two examples of different designs implying different AEEDs in Table 2, which are both also different from the original AEED in Table 1a.

While our original, individualistic exposure mapping $d_i = g_i(\mathbf{z}) = z_i$ does not satisfy SUTVA, consider the exposure mapping, $d_i = g_i(\mathbf{z}) = z_{3-i}$, which parallels the true decision rule (i.e., an individual is exposed if and only if their housemate is treated). Now the AEED is 1, regardless of what probabilities we assign to each feasible assignment (provided the relevant individual positivity conditions hold; illustrated in Table 3). Informally, SUTVA has been "restored" with respect to the exposure and we no longer have issues in characterizing the causal effect of exposing a unit on its outcome.

Whether each stability assumption holds depends on the outcome of interest, how exposures are defined, the design used to induce exposures, and the real, underlying structure of the outcome generating process. For example, consider a setting where there is partial interference on a given outcome. That is, there may be arbitrary interference within, but not across mutually exclusive clusters of units. In such a setting, NURVA will hold with respect to that outcome on all designs where all units in a cluster are assigned to a single exposure level, and the AEED will be equivalent to the global average treatment effect (the effect of treating all subjects relative to no subjects). However, NURVA will not necessarily hold for designs that result in within-cluster variation in treatment assignment.[9] We address considerations for cross-design generalization further in Section 6.

Because NURVA embeds all observable implications of SUTVA, the appropriate interpretation in a given setting should be determined by the researcher. The structure of $g(\cdot)$ is a modeling choice rather

[9] In cluster-randomized designs with partial interference, SUTVA is stronger: it would also require invariance to off-support within-cluster reallocations, implying that, for example, response for an observation when it is uniquely treated within a cluster of otherwise untreated observations is identical to response for that observation when all units in the cluster are treated.

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

Table 2. Different designs yield different AEEDs

(a) Alternate design 1				
z	**d**	$y_1(\mathbf{z})$	$y_2(\mathbf{z})$	$Pr[\mathbf{Z} = \mathbf{z}]$
(0,0)	(0,0)	0	0	0.5
(0,1)	(0,1)	1	0	0
(1,0)	(1,0)	0	1	0
(1,1)	(1,1)	1	1	0.5
$\longrightarrow \tau(1,0) = 1$				
(b) Alternate design 2				
z	**d**	$y_1(\mathbf{z})$	$y_2(\mathbf{z})$	$Pr[\mathbf{Z} = \mathbf{z}]$
(0,0)	(0,0)	0	0	0.25
(0,1)	(0,1)	1	0	0.25
(1,0)	(1,0)	0	1	0.25
(1,1)	(1,1)	1	1	0.25
$\longrightarrow \tau(1,0) = 0$				

Table 3. Exposure mapping that satisfies SUTVA

(a) Alternate design 1				
z	**d**	$y_1(\mathbf{z})$	$y_2(\mathbf{z})$	$Pr[\mathbf{Z} = \mathbf{z}]$
(0,0)	(0,0)	0	0	0.5
(0,1)	(1,0)	1	0	0
(1,0)	(0,1)	0	1	0
(1,1)	(1,1)	1	1	0.5
$\longrightarrow \tau(1,0) = 1$				
(b) Alternate design 2				
z	**d**	$y_1(\mathbf{z})$	$y_2(\mathbf{z})$	$Pr[\mathbf{Z} = \mathbf{z}]$
(0,0)	(0,0)	0	0	0.25
(0,1)	(1,0)	1	0	0.25
(1,0)	(0,1)	0	1	0.25
(1,1)	(1,1)	1	1	0.25
$\longrightarrow \tau(1,0) = 1$				

than a direct implication of the design and may be based on theory or other knowledge of the underlying data-generating processes. We also note that "coarsened" exposures—which do not represent unique potential outcomes—induce design-dependent version weights. If the researcher is concerned about hidden versions within a potentially coarse exposure, they may refine g to include the plausible versions, or they may report features of the induced version (e.g., the distribution of treated-neighbor counts among "indirectly exposed"). Additionally, when comparing across studies, researchers may wish to standardize to a common target version distribution—if such version-specific effects are estimable. To gain prior information about interference structure itself, researchers might use design and analytic recommendations in, for example, Bowers *et al.* (2018) and Baird *et al.* (2018).

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

Practically, the NURVA/SUTVA distinction changes not the estimator used within a study, but the claim attached to it. In the rebel-group survey, the analyst must specify the implemented assignment mechanism and its support (e.g., surveying 3 of 10 groups in random order), choose and justify $g(\cdot)$ (e.g., "surveyed"/"not" versus order- or enumerator-specific exposures) and state the set of feasible interventions relevant to the substantive question. Invoking NURVA means that $\tau(d,d')$ is interpreted as an effect of the researcher-defined exposure under the realized protocol, on $\text{Supp}(\mathbf{Z})$. Invoking SUTVA adds the stronger claim that the same contrast is invariant to off-support allocations in the feasible set, a claim not implied by the design alone. Accordingly, applied reports should make $\mathbf{Z}$, $\text{Supp}(\mathbf{Z})$, $g(\cdot)$, and inferential scope explicit, and, when feasible, probe fragility by refining $g(\cdot)$ within the design.

6. A generalizing assumption

SUTVA net of NURVA is a *generalizing assumption*, in that it speaks to causal effects outside of the experiment. We also often wish to use the results from our population to learn about the effect of treating a larger population, a different population or even the same population at a different time. Much of the modern work on external validity takes SUTVA for granted and treats treatment effect heterogeneity as the primary obstacle to generalizability. We provide another perspective. First, we consider a more narrow sense of generalization, as we hold fixed the experimental population and consider alternative designs that intervene on this population. Then we show that even with NURVA and internally identified AEEDs, inferring to alternative populations requires additional assumptions about the alignment of exposure versions and their design-induced distributions.

6.1. Generalizing across designs: General equilibrium

It is sometimes the case that we are interested in the effect of treating everyone versus treating no one, but what we have in practice is a design where some units are randomly assigned to each condition. Table 4 constructs two schedules of raw potential outcomes. In both cases, the AEED from the observed design, where $\text{Supp}(\mathbf{Z}) = \{(1,0),(0,1)\}$, does not reflect the AEED in which treatment is delivered uniformly over the population, where $\text{Supp}(\mathbf{Z}') = \{(1,1),(0,0)\}$.

These schedules capture mechanisms in which the effect of "being treated" changes as coverage moves from partial to universal—for example, job training that is advantageous only when few workers are trained or vaccinations whose effectiveness depends on population coverage reaching a critical threshold. The instability of AEEDs for this limited notion of generalization across designs with a fixed set of experimental units raises concerns for external validity—generalization to units outside our original experimental sample, as the term is commonly used in political science. The limitations of external validity under NURVA alone can be seen by expanding the table of raw potential outcomes with new units. Alternatively, we can explicitly embed a sampling–treatment design under this framework, which we do below.

6.2. Hidden treatment variations

Both SUTVA and NURVA require an assumption that potential outcomes are stable under the researcher-specified exposures but NURVA requires this stability only on the support of the design, while SUTVA requires it under all feasible designs. When SUTVA is violated, even if NURVA holds, design-specific AEEDs need not generalize to other designs.

To illustrate this, consider an experiment where the design space is $\mathcal{Z} \subseteq \{0,1,2\}^N$, and the exposure mapping is

$$g_i(\mathbf{z}) = \begin{cases} 1 & : z_i > 0 \\ 0 & : \text{otherwise} \end{cases}, \forall i \in \{1,\ldots,N\}, \mathbf{z} \in \{0,1,2\}^N.$$

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

Table 4. General equilibria

(a) Attenuation under uniform rollout					
$\mathbf{z}$	$\mathbf{d}$	$y_1(\mathbf{z})$	$y_2(\mathbf{z})$	$Pr[\mathbf{Z}=\mathbf{z}]$	$\Pr[\mathbf{Z}'=\mathbf{z}]$
(0,0)	(0,0)	0.5	0.5	0	0.5
(0,1)	(0,1)	0	1	0.5	0
(1,0)	(1,0)	1	0	0.5	0
(1,1)	(1,1)	0.5	0.5	0	0.5
$\longrightarrow \tau_{\mathbf{Z}}(1,0) = 1$					
$\longrightarrow \tau_{\mathbf{Z}'}(1,0) = 0$					
(b) Amplification under uniform rollout					
$\mathbf{z}$	$\mathbf{d}$	$y_1(\mathbf{z})$	$y_2(\mathbf{z})$	$Pr[\mathbf{Z}=\mathbf{z}]$	$\Pr[\mathbf{Z}'=\mathbf{z}]$
(0,0)	(0,0)	0	0	0	0.5
(0,1)	(0,1)	0	0	0.5	0
(1,0)	(1,0)	0	0	0.5	0
(1,1)	(1,1)	1	1	0	0.5
$\longrightarrow \tau_{\mathbf{Z}}(1,0) = 0$					
$\longrightarrow \tau_{\mathbf{Z}'}(1,0) = 1$					

Table 5. Hidden treatment variations

(a) Design $\mathbf{Z}$: assigns only in $\{0,1\}$				
$\mathbf{z}$	$\mathbf{d}$	$y_1(\mathbf{z})$	$y_2(\mathbf{z})$	$\Pr[\mathbf{Z}=\mathbf{z}]$
(0,1)	(0,1)	0	1	0.5
(1,0)	(1,0)	1	0	0.5
$\longrightarrow \tau_{\mathbf{Z}}(1,0) = 1$				
(b) Design $\mathbf{Z}'$: assigns only in $\{0,2\}$				
$\mathbf{z}$	$\mathbf{d}$	$y_1(\mathbf{z})$	$y_2(\mathbf{z})$	$\Pr[\mathbf{Z}'=\mathbf{z}]$
(0,2)	(0,1)	0	2	0.5
(2,0)	(1,0)	2	0	0.5
$\longrightarrow \tau_{\mathbf{Z}'}(1,0) = 2$				

For this example, assume that there is no interference in the sense that $y_i(\mathbf{z}) = y_i(\mathbf{z}')$ if $z_i = z_i'$. This allows us to focus on the implications of hidden treatment variations in isolation. If treatments 1 and 2 are associated with different potential outcomes, then the AEED could vary between designs that do and do not sample treatment 2, as in Table 5. (For simplicity, Table 5 displays only assignments in $\text{Supp}(\mathbf{Z})$ and $\text{Supp}(\mathbf{Z}')$; all other feasible assignments have probability zero under the respective designs.)

In the rebel group survey, different question wordings or enumerator identities may elicit different reports, yet all such surveys are coded as "surveyed." In the school experiment, untreated students with one or several treated friends are both coded as "indirectly treated." When these versions are associated with different potential outcomes, the AEED under a given design depends on how that design weights these version-specific potential outcomes.

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

6.3. Generalizing across sampling–treatment designs

We now sketch how the framework extends to joint sampling–treatment designs, linking causal and sampling inference. Within a finite population of size N, consider a design $\mathbf{Z}$ that first draws a simple random sample of size n without replacement and then, among sampled units, assigns k to treatment and $n-k$ to control by complete randomization. Define an exposure mapping where $z_i = 2$ indicates that unit i is sampled and treated, $z_i = 1$ that it is sampled and untreated and $z_i = 0$ that it is not sampled. We can define the AEED, $\tau_{\mathbf{Z}}(2,1)$ which, under NURVA, admits a causal interpretation as the effect, under this joint sampling-assignment design, of exposing a unit to being "sampled and treated" rather than "sampled and untreated."

Now consider a different design $\mathbf{Z}'$ on the same population, with sample size n' and number treated k'. Using the same exposure mapping, define the corresponding AEED $\tau_{Z'}(2,1)$.[10]

We define the exposure mapping as

$$g_i(z) = \begin{cases} 2 & : z_i = 2 \\ 1 & : z_i = 1 \\ 0 & : z_i = 0 \end{cases}, \forall i \in \{1,\ldots,N\}, \mathbf{z} \in \mathcal{Z}.$$

Even if NURVA holds for both $\mathbf{Z}$ and $\mathbf{Z}'$, there is no guarantee that $\tau_{\mathbf{Z}}(2,1)$ will equal $\tau_{\mathbf{Z}'}(2,1)$: changing either n to n' or k to k' changes the distribution of other units' sampling or treatment statuses conditional on $D_i = d$, and hence may change the EPOs $y_i(d)$ that enter the AEPOs. Learning what would happen under alternative values of (n,k) or in alternative target populations requires additional stability assumptions on how $y_i(d)$ varies with the sampling and treatment scheme, as in the broader generalizability and transportability literature. For example, Tipton (2013) imposes dual SUTVA requirements for both treatment assignment (SUTVA($\mathcal{S}$)) and sample selection (SUTVA($\mathcal{P}$)), and Lesko *et al.* (2017) assume that the distribution of treatment versions and, under interference, the pattern of interference are the same in the study sample and the target population. Because assignments in $\mathcal{Z} \setminus \text{Supp}(\mathbf{Z})$ are never observed under a single fixed design, off-support invariance is not nonparametrically testable within that design. Evidence about SUTVA must therefore come from additional designs or auxiliary structure that place competing versions on support—for example, saturation designs that vary treated fractions, multi-version interventions, repeated implementations that vary protocol features, such as wording or enumerators, or observational settings with known assignment probabilities over a wider support. Such evidence can make off-support invariance more or less credible, but it is an additional empirical argument, not a consequence of the original design.

7. Estimation and inference

Estimation and inference are possible even when SUTVA does not hold, with NURVA providing enough structure to achieve traditional results. In particular, NURVA implies that the statistical results from the case where SUTVA holds can be adapted, given the statistical and observational equivalence of SUTVA and NURVA for a given design. We do not provide details on these results in the main text; the Supplementary Material details the theory of estimation and inference. We focus on the use of inverse probability weighted estimators for estimation of the AEPO and the AEED and describe properties of the estimator in this setting, which largely follow from results on Aronow and Samii (2017). As in prior

[10] Support under the two designs:

$$\text{Supp}(\mathbf{Z}) = \text{Sym}\big((\underbrace{2,\ldots,2}_{k},\underbrace{1,\ldots,1}_{n-k},\underbrace{0,\ldots,0}_{N-n})\big)$$
$$\text{Supp}(\mathbf{Z}') = \text{Sym}\big((\underbrace{2,\ldots,2}_{k'},\underbrace{1,\ldots,1}_{n'-k'},\underbrace{0,\ldots,0}_{N-n'})\big).$$

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

work, we demonstrate finite-sample unbiasedness of Horvitz–Thompson-type estimators (both with and without covariate adjustment), provide a conservative variance estimator and discuss asymptotic results under appropriate regularity conditions. In a result we believe to be novel, we also show that finite-sample unbiasedness can be attained for covariate-adjusted estimators even if NURVA does not hold. Supporting derivations and regularity conditions are included in the Supplementary Material.

8. Conclusion

We contribute to the growing literature on the design-based approach by reexamining the assumptions underpinning the design-based framework. We propose as estimands the AEPO and the AEED; counterparts to conventional estimands that arise in this framework when we refrain from imposing assumptions such as SUTVA on the data-generating process. These quantities remain well-defined under arbitrary interference or exposure misspecification: absent stability assumptions, they represent design-conditional mixtures and mixture contrasts, not generally causal effects of exposures. These design-specific quantities are useful when the implemented protocol itself is the object of inference—for example, a fixed rollout rule, survey procedure, saturation design or network intervention with only some feasible allocations—even when broader off-support invariance is not credible. What NURVA buys in such cases is a within-support causal interpretation; what it does not buy is extrapolation to alternative allocations, protocols or populations. SUTVA is required for those off-support interpretations. Throughout, we take the exposure mapping $g(\cdot)$ as a modeling choice that must be justified rather than assumed away. Different exposure mappings encode different substantive views about which versions of treatment and which channels of interference matter, and if the mapping is badly mismatched to the data-generating process, the resulting AEPOs and AEEDs may fail to align with the scientific questions of interest. For applied researchers, the practical implication is to specify $\mathcal{Z}, \mathbf{Z}, \text{Supp}(\mathbf{Z}), g(\cdot)$, and inferential scope explicitly, state whether the identifying claim is no stability assumption, NURVA, or SUTVA, and interpret the estimand accordingly. This reframing makes explicit the quantities that researchers may recover in the absence of SUTVA and the additional claims needed to generalize beyond the implemented design.

We provide a conservative variance estimator under NURVA in the Supplementary Material. There is additional hope for further refining inference on targets such as AEEDs from Sävje (2024); inference on AEEDs is asymptotically equivalent to inference on the "average direct effect" $N^{-1}\sum_{i=1}^{N}(\mathrm{E}[y_i(d;\mathbf{D}_{-i}) - y_i(d';\mathbf{D}_{-i})])$ if both statistical dependence in $\mathbf{D}$ and unmodeled interference are sufficiently local. Extensions to observational settings, or settings where treatment assignment or sampling procedures are known but the known design is somehow "broken" by, for example, missingness correlated with response, would require estimation of these probabilities, in the vein of Robins, Rotnitzky, and Zhao (1994) and Hirano, Imbens, and Ridder (2003). Moreover, inference may be possible under weaker assumptions than NURVA, such as factorial experiments, when there are multiple assignments by treating one assignment as random and holding the rest of the assignments as fixed. We leave such extensions to future work.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/pan.2026.10045.

Acknowledgements. Thank you to Kirk Bansak, Haoge Chang, Issa Kohler-Hausmann, Jens Hainmueller, Fredrik Sävje and Cyrus Samii for their insightful comments and discussions on this topic. The authors used OpenAI's GPT-5.5 Pro to assist with language editing and proofreading during proof correction. The authors reviewed all AI-assisted suggestions and are solely responsible for the article's content.

References

Aronow, P. M., and J. A. Middleton. 2013. "A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments." *Journal of Causal Inference* 1 (1): 135–154.

Aronow, P. M., and C. Samii. 2017. "Estimating Average Causal Effects under General Interference, with Application to a Social Network Experiment." *The Annals of Applied Statistics* 11 (4): 1912–1947.

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

Baird, S., J. Aislinn Bohren, C. McIntosh, and B. Özler. 2018. "Optimal Design of Experiments in the Presence of Interference." *Review of Economics and Statistics* 100 (5): 844–860.
Basse, G. W., A. Feller, and P. Toulis. 2019. "Randomization Tests of Causal Effects under Interference." *Biometrika* 106 (2): 487–494.
Borusyak, K., P. Hull, and X. Jaravel. 2025. "Design-Based Identification with Formula Instruments: A Review." *The Econometrics Journal* 28 (1): 83–108.
Bowers, J., B. A. Desmarais, M. Frederickson, N. Ichino, H.-W. Lee, and S. Wang. 2018. "Models, Methods and Network Topology: Experimental Design for the Study of Interference." *Social Networks* 54: 196–208.
Card, D. 2022. "Design-Based Research in Empirical Microeconomics." *American Economic Review* 112 (6): 1773–1781.
Centola, D. 2010. "The Spread of Behavior in an Online Social Network Experiment." *Science* 329 (5996): 1194–1197.
Cole, S. R., and M. A. Hernán. 2008. "Constructing Inverse Probability Weights for Marginal Structural Models." *American Journal of Epidemiology* 168 (6): 656–664.
Cox, D. R. 1958. *Planning of Experiments*. New York: Wiley.
Dunning, T. 2010. "Design-Based Inference: Beyond the Pitfalls of Regression Analysis?" In D. Collier & H. Brady (Eds.), *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 273–311. Lanham, MD: Rowman & Littlefield.
Gao, M., and P. Ding. 2025. "Causal Inference in Network Experiments: Regression-Based Analysis and Design-Based Properties." *Journal of Econometrics* 252: 106119.
Harshaw, C., Y. Wang, and F. Sävje. 2022. "A Design-Based Riesz Representation Framework for Randomized Experiments." In *NeurIPS 2022 Workshop on Causality for Real-World Impact*. https://openreview.net/forum?id=GM16YCLkEmz.
Heng, S., J. Zhang, and Y. Feng. 2025. "Design-Based Causal Inference with Missing Outcomes: Missingness Mechanisms, Imputation-Assisted Randomization Tests, and Covariate Adjustment." *Journal of the American Statistical Association* 121 (553): 312–325. Just-accepted.
Hernán, M. A., and J. M. Robins. 2006. "Estimating Causal Effects from Epidemiological Data." *Journal of Epidemiology & Community Health* 60 (7): 578–586.
Hirano, K., G. W. Imbens, and G. Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71 (4): 1161–1189.
Hudgens, M. G., and M. E. Halloran. 2008. "Toward Causal Inference with Interference." *Journal of the American Statistical Association* 103 (482): 832–842.
Kalla, J. L., and D. E. Broockman. 2016. "Campaign Contributions Facilitate Access to Congressional Officials: A Randomized Field Experiment." *American Journal of Political Science* 60 (3): 545–558.
Kish, L. 1965. *Survey Sampling*. New York: Wiley.
Kocher, M. A., and N. P. Monteiro. 2016. "Lines of Demarcation: Causation, Design-Based Inference, and Historical Research." *Perspectives on Politics* 14 (4): 952–975.
Lesko, C. R., A. L. Buchanan, D. Westreich, J. K. Edwards, M. G. Hudgens, and S. R. Cole. 2017. "Generalizing Study Results: A Potential Outcomes Perspective." *Epidemiology* 28 (4): 553–561.
Leung, M. P. 2020. "Treatment and Spillover Effects under Network Interference." *Review of Economics and Statistics* 102 (2): 368–380.
Lu, S., L. Shi, Y. Fang, W. Zhang, and P. Ding 2025. "Design-Based Causal Inference in Bipartite Experiments." Preprint, arXiv:2501.09844.
Manski, C. F. 2013. "Identification of Treatment Response with Social Interactions." *The Econometrics Journal* 16 (1): S1–S23.
Middleton, J. A. 2018. "A Unified Theory of Regression Adjustment for Design-Based Inference." arXiv: 1803.06011 [math.ST]. https://arxiv.org/abs/1803.06011
Miguel, E., and M. Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1): 159–217.
Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625.
Neyman, J. 1990 [1923]. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9 (Reprint)." *Statistical Science* 5 (4): 465–472.
Neyman, J., K. Iwaszkiewicz, and S. Kolodziejczyk. 1935. "Statistical Problems in Agricultural Experimentation." *Supplement to the Journal of the Royal Statistical Society* 2 (2): 107–180.
Robins, J. M., A. Rotnitzky, and L. P. Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–866.
Rubin, D. B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.
Särndal, C.-E. 1978. "Design-Based and Model-Based Inference in Survey Sampling." *Scandinavian Journal of Statistics* 5 (1): 27–43.
Sävje, F. 2024. "Causal Inference with Misspecified Exposure Mappings: Separating Definitions and Assumptions." *Biometrika* 111 (1): 1–15.
Sävje, F., P. M. Aronow, and M. Hudgens. 2021. "Average Treatment Effects in the Presence of Unknown Interference." *The Annals of Statistics* 49 (2): 673–701.

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

Sawa, T. 1978. "Information Criteria for Discriminating among Alternative Regression Models." *Econometrica* 46 (6): 1273–1291.

Sinclair, B., M. McConnell, and D. P. Green. 2012. "Detecting Spillover Effects: Design and Analysis of Multilevel Experiments." *American Journal of Political Science* 56 (4): 1055–1069.

Tipton, E. 2013. "Improving Generalizations from Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts." *Journal of Educational and Behavioral Statistics* 38 (3): 239–266.

Wang, Y., and M. Jetsupphasuk. 2021. "Causal Inference in Longitudinal Data Under Unknown Interference." Preprint, arXiv:2106.15074.

Zhao, A., and P. Ding. 2022. "Reconciling Design-Based and Model-Based Causal Inferences for Split-Plot Experiments." *The Annals of Statistics* 50 (2): 1170–1192.

https://doi.org/10.1017/pan.2026.10045 Published online by Cambridge University Press

Cite this article: Aronow, P. M., Jang, A. and Offer-Westort, M. (2026). On the Foundations of the Design-Based Approach. *Political Analysis*, 1–16. https://doi.org/10.1017/pan.2026.10045