Designing Transportable Experiments Under S-admissability

My Phan
University of Massachusetts Amherst

David Arbour*
Adobe Research

Drew Dimmery*
Facebook Core Data Science

Anup B. Rao*
Adobe Research

Abstract

We consider the problem of designing a randomized experiment on a source population to estimate the Average Treatment Effect (ATE) on a target population. We propose a novel approach which explicitly considers the target when designing the experiment on the source. Under the covariate shift assumption, we design an unbiased importance-weighted estimator for the target population's ATE. To reduce the variance of our estimator, we design a covariate balance condition (Target Balance) between the treatment and control groups based on the target population. We show that Target Balance achieves a higher variance reduction asymptotically than methods that do not consider the target population during the design phase. Our experiments illustrate that Target Balance reduces the variance even for small sample sizes.

1 Introduction

The problem of generalization is present everywhere that experiments are run. In the online environment, tests are run with the users who show up on the product while the experiment is running (and are therefore highly active users), while inferences about user experience are most useful on the full set of users (both highly active and less active) [Wang et al., 2019]. In clinical research, it is an omnipresent problem to recruit minorities into randomized trials [Fisher and Kalbaugh, 2011], thus making it difficult to assume that the measured effects will generalize to the larger population of interest (e.g. the United States as a whole, or people afflicted with a particular health condition). In lab experiments, the sample is often one of convenience such as undergraduates in rich countries or from pools of potential subjects available online [Henrich et al., 2010]. Field experiments in governance or development such as Dunning et al. [2019] are conducted in particular countries or in particular communities, but the policy implications of such work stretch far beyond the borders of the study population. As in Dunning et al. [2019], the desire is not just to understand how Burkina Faso voters respond to more information about their political leaders, but to understand how voters across the world might respond to similar informational treatments. The same is true for experiments in development economics, such as microfinance [Meager, 2019] and for studies of internet phenomena [Munger, 2018]. In these cases, it isn't a *surprise* after running an experiment that generalizing the knowledge is important; indeed, generalization of knowledge to a broader population is core to the motivation for the experiment in the first place.

We pre-suppose that an experimenter knows ex-ante the population on which they wish to draw broader inferences. The task we consider, therefore, is to design an experiment that best allows the generation of causal knowledge on this inferential target. While previous work [Hartman et al., 2015, Dehejia et al., 2019, Stuart et al., 2011, DuGoff et al., 2014] has examined corrections on the analysis side to extrapolate estimates from sample to target population, the novelty of this work is in doing this through a *design-based* solution. That is, if you know your goal is to generalize to a target population, we consider how that should modify experimental design. We focus in particular on the "S-admissability" condition for transportability, in which the outcome distribution conditional on a set of covariates is the same in both the source and target distributions [Pearl and Bareinboim, 2011]. **Contributions.** Using the Mahalanobis distance and importance weighting, we design an estimator with a balancing condition for the target distribution's ATE that is unbiased and has low variance.

- In Section 3, we introduce an importance-weighted

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

*Equal contribution.

estimator with a balance condition called Target Balance that explicitly considers the target distribution in the design phase.
- In Section 5.1 we show that using the importance-weighted estimator with Target Balance results in an unbiased estimator of the target distribution's ATE (Theorem 1)
- We analyze the variance assuming a linear model. In Section 5.2.1, we show that when the dimension of the covariates $d = 1$, for a finite sample size n, Target Balance reduces the variance (Corollary 1). Moreover, among all balance criteria with rejection probability at most α (including balancing by only considering the source distribution, which we call Source Balance), Target Balance achieves the optimal variance reduction (Theorem 2). When $d \geq 1$ (Section 5.2.2), when the sample size is large, Target Balance reduces the variance (Theorem 3) and achieves a lower variance than Source Balance (Theorem 4).
- In Section 6 we perform experiments[1] to show that Target Balance has small mean-squared errors even for $d > 1$, small sample size and non-linear model.

2 Problem Setting

We first fix notation before proceeding to the problem setting. Upper-case letters are used to denote random variables, lower-case letters are used to denote values taken by them. We use bold-faced letters to denote n samples and normal letters to denote a single sample. For example, $X_i \in \mathcal{R}^d$ is a random variable denoting the covariates of sample i. $\mathbf{X} = (X_1, ..., X_n)^T \in \mathcal{R}^{n \times d}$ is the random variables $X_1, .., X_n$ concatenated together. $x_i \in \mathcal{R}^d$ is a value of X_i, and $\mathbf{x} = (x_1, ..., x_n)^T \in \mathcal{R}^{n \times d}$ is a value of $\mathbf{X}$.

Some random variables, like X, can have two different distributions, either source distribution or target distribution. In that case, we use $\mathbb{E}^S X$, $\text{var}^S X$ and $cov^S X$ to denote the expectation, variance and covariance with respect to the source distribution, and $\mathbb{E}^T X$, $\text{var}^T X$ and $cov^T X$ to denote with respect to the target distribution. We use no superscripts when there is no confusion. For example $\mathbb{E}A_i$ is the expectation of the treatment assignment A_i of sample i. For a random variable R, we use $\mathbb{E}_R$, var_R and cov_R to denote the expectation, variance and covariance over the randomness of R. For example, $\mathbb{E}^S_{\mathbf{X}}$ denote the expectation over the randomness of $\mathbf{X} = (X_1, ..., X_n)^T$ according to the source distribution. We omit the subscripts when it's clear.

The problem considered in this paper is as follows. We assume that we are presented with two populations, referred to as the source and target populations, with corresponding densities p_S and p_T, respectively. We further assume that we observe a set of pre-treatment covariates from the source population, $x_1, \ldots x_n \sim p_S$. We assume that we are freely able to assign treatment, $a_1, \ldots, a_n \in \{0, 1\}$ to individuals observed in the source population and observe their outcomes, $y_1, \ldots, y_n \in \mathcal{R}$. The estimand of interest is the average treatment effect for the *target* population (the population of individuals which were not subject to an experiment),

$$\tau^T_Y = \mathbb{E}^T[Y^{A=1} - Y^{A=0}]. \tag{1}$$

Where $Y^{A=0}$, $Y^{A=1}$ are the potential outcomes [Rubin, 2011], i.e., the values of Y that would have been observed had treatment been observed at $A = 1$ or $A = 0$, respectively. We use Y to denote (Y^0, Y^1) and Y^* to denote the *observed* outcome.

In order to make this problem tractable we will assume the following throughout the remainder of the paper:

Assumption 1. *Equality of conditional densities, i.e., $p_S(Y|X) = p_T(Y|X)$ (note $p_S(X) \neq p_T(X)$ in general).*

This assumption places identification of the transportability of effects under the rubric of S-admissability [Pearl and Bareinboim, 2011].

Assumption 2. *Overlap between source and target distributions, i.e., $p_T(X) > 0 \implies p_S(X) > 0$.*

Assumption 3.

$$Y^1 = \psi(X)^T \beta_1 + \mathcal{E}_1 \qquad Y^0 = \psi(X)^T \beta_0 + \mathcal{E}_0$$

where ψ is a basis function and $\mathcal{E}_1, \mathcal{E}_0$ are mean zero random variables.

To reduce notational clutter, and without loss of generality, we will assume that ψ is the identity function for the remainder of paper so that we can write X instead of $\psi(X)$.

Assumption 4. *The ratio of the pdfs, $p_T(X)/p_S(X)$, is known.*

In a nested trial design in which the sampling probabilities from the population are known [Dahabreh et al., 2019], this ratio can be calculated as

$$\frac{p_T(X)}{p_S(X)} = \frac{p(X|S=0)}{p(X|S=1)} = \frac{p(S=0|X)}{p(S=1|X)} \frac{p(S=1)}{p(S=0)},$$

where $S = 1$ indicates that the unit is selected to be in the source and $S = 0$ indicates that the unit is not selected and is in the target.

[1] Code for this paper is available at `https://github.com/myphancs/Designing_Transportable_Experiments`

The assumptions, though nontrivial, are common throughout the literature on transportability [Stuart et al., 2011, Hartman et al., 2015, Pearl and Bareinboim, 2011]. We conjecture that similar results to those in this paper will hold in the case in which importance weights are estimated with parametric convergence rates. We leave this extension as future work.

For a sample i, let X_i, A_i, Y_i^a and Y_i^* be the covariates, treatment, outcome of treatment a, and observed outcome. Let n_0 be the size of the control group (where $A_i = 0$) and n_1 be the size of the treatment group (where $A_i = 1$). Similar to common practice (c.f., Stuart et al. [2011], Hartman et al. [2015], Rudolph and van der Laan [2017], Buchanan et al. [2018]), we infer τ_Y^T with importance weights,

$$\begin{aligned}\hat{\tau}_Y^T &:= \frac{1}{n_1}\sum_{A_i=1} W_i Y_i^* - \frac{1}{n_0}\sum_{A_i=0}^{n} W_i Y_i^* \\ &= \frac{1}{n_1}\sum_{i=1}^{n} W_i A_i Y_i^1 - \frac{1}{n_0}\sum_{i=1}^{n} W_i(1-A_i)Y_i^0 \qquad (2)\end{aligned}$$

where $W_i = \frac{p_T(X_i)}{p_S(X_i)}$. While equation 2 is unbiased, the estimate can incur large variance in the presence of large importance weights.

For ease of notation we define $Z_i = 2A_i - 1 \in \{-1, 1\}$ and let $\mathbf{Z}$ be the $n \times 1$ vector of random variables $Z_1, ..., Z_n$ and $\mathbf{z}$ be a value taken by $\mathbf{Z}$. $Y_i = (Y_i^0, Y_i^1)$ is a random variable denoting all possible outcomes of sample i. $\mathbf{Y} = (Y_1, ..., Y_n)^T \in \mathcal{R}^{n\times 2}$ is the random variables $Y_1, .., Y_n$ concatenated together. $y_i \in \mathcal{R}^2$ is a value of Y_i, and $\mathbf{y} = (y_1, ..., y_n)^T \in \mathcal{R}^{n\times 2}$ is a value of $\mathbf{Y}$. Let $w_i = \frac{p_T(x_i)}{p_S(x_i)}$ and $\mathbf{w}$ be the $n \times n$ diagonal matrix with $\mathbf{w}(i, i) = w_i$. For matrix $\mathbf{a}$, we use $\tilde{\mathbf{a}}$ to denote $\mathbf{wa}$ where each row i of $\mathbf{a}$ is multiplied by w_i.

3 Designing for Transportation

We consider $n_0 = n_1 = n/2$ throughout the paper. The core contribution of this work is a procedure to estimate equation 1 which explicitly considers the target population when designing the experiment for the source population. We focus on adapting re-randomization, an experimental design procedure which optimizes balance, i.e., the difference in means of X between treatment and control groups. Specifically, rerandomization centers on a *balance criterion*,

Definition 1 (Target Balance)**.** *With a rejection threshold α, define the balance condition:*

$$\phi_T^\alpha(\mathbf{x}, \mathbf{Z}) = \begin{cases} 1, & \text{if } M(\frac{2}{n}(\mathbf{wx})^T\mathbf{Z}) < a(\mathbf{x}) \\ 0, & \text{otherwise} \end{cases}$$

where $M(\frac{2}{n}(\mathbf{wx})^T\mathbf{Z})$ is a distance (defined below in Eq. 3) between the covariates associated with treatment and control given by $\mathbf{Z}$ and $a(\mathbf{x})$ is chosen such that $\mathbb{P}(\phi_T^\alpha = 1|\mathbf{x}) = 1 - \alpha$.

We omit α and simply write ϕ_T when α is not necessary for exposition. We omit $\mathbf{x}$ and write a when there is no confusion.

The full assignment procedure is then

1. Assign A randomly for each person $1, \ldots, n$ such that $\sum_i a_i = n_1$. There are $\binom{n}{n_1}$ ways to choose this, each of which is equally likely.
2. If $\phi_T(\mathbf{x}, \mathbf{z}) = 0$ return to step (1).
3. Conduct experiment with treatment assignments, A.

Following standard practice in rerandomization [Morgan et al., 2012], we will focus on a criterion based on Mahalanobis distance, but incorporating a weighting term to express our desire for balance in the *target* distribution rather than in the source. We refer to this weighted Mahalanobis distance as $M(\frac{2}{n}(\mathbf{wx})^T\mathbf{Z})$, where $M(\cdot)$ is defined as:

$$\begin{aligned}M(U) &:= (U)^T\, Cov(U)^{-1}\, (U) \\ &= \|B\|^2 \text{ where } B = UCov(U)^{-1/2} \qquad (3)\end{aligned}$$

Thus, the balance condition $M(\frac{2}{n}(\mathbf{wx})^T\mathbf{Z}) < a$ is equivalent to truncating the square norm of B to be less than a. Note that this is a standardized measure of the difference in importance-weighted covariate-means between treatment and control, since

$$\frac{2}{n}\sum_{Z_i=1} w_i x_i - \frac{2}{n}\sum_{Z_i=-1} w_i x_i = \frac{2}{n}(\mathbf{wx})^T\mathbf{Z}.$$

Thus, rerandomization simply rejects designs with covariate imbalance larger than a pre-specified value.

The novelty in our proposed design is to reject samples based on imbalance in the the *target* distribution rather than based on imbalance in the *source* distribution. The standard in the rerandomization literature is to focus on balance in the *source* distribution, which in our setup implies assuming that the target distribution is equal to the source distribution. Therefore, importance weights in this case are all equal to one. We call this balancing condition *Source Balance*, which we denote by $\phi_S^\alpha(\mathbf{x}, \mathbf{Z})$.

We explain the intuition behind using Target Balance rather than Source Balance. An unbiased estimator for the source's ATE is: $\hat{\tau}_Y^S := \frac{1}{n_1}\sum_i A_i Y_i^1 - \frac{1}{n_0}\sum_i A_i Y_i^0$. There are existing results [Li et al., 2018, Harshaw et al., 2019] that can be applied to linear models to show variance reduction of $\hat{\tau}_Y^S$ with Source Balance defined by $\mathbf{x}$.

By defining new variables $\tilde{Y}_i^a = W_i \cdot Y_i^a$ for $a \in \{0, 1\}$, the importance weighted estimator can now be ex-

pressed in a similar form to $\hat{\tau}_Y^S$: $\hat{\tau}_Y^T = \frac{1}{n_1}\sum_i A_i\tilde{Y}_i^1 - \frac{1}{n_0}\sum_i A_i\tilde{Y}_i^0$.

We are no longer in a linear setting because $\tilde{Y}_i^a = W_i(\beta_a^T X_i + \mathcal{E}_a)$. But by defining $\tilde{X}_i = W_i X_i$ and $\tilde{\mathcal{E}} = W_i\mathcal{E}$ we have: $\tilde{Y}_i^a = \beta_a^T\tilde{X}_i + \tilde{\mathcal{E}}_a$, where $\mathbb{E}\left[\tilde{Y}^a|X\right]$ is a linear function of $\tilde{X}$, which can be considered as a feature-transformed X. [Li et al., 2018, Harshaw et al., 2019] can now be applied to estimate $\hat{\tau}_Y^T$ with Target Balance defined by $\tilde{\mathbf{x}} = \mathbf{wx}$.

Let $\rho(\mathbf{x}, \mathbf{z}) \in \{0, 1\}$ be a function of $\mathbf{x}$ and $\mathbf{z}$ used in the re-randomization procedure. Note that since we re-sample $\mathbf{Z}$ until $\rho = 1$, only the distribution of $\mathbf{Z}$ is affected by the balance condition. Let $\mathbf{Z}_\rho$ denote the distribution of $\mathbf{Z}$ after being accepted by the balance condition $\rho = 1$.

4 Related Work

Our work relates to and ties together two distinct strands of research: (1) ex-post generalization of experimental results to population average effects and (2) ex-ante experimental design. We will discuss each in turn.

Generalization.

Within the literature on methods for generalization, work has generally focused on ex-post adjustments to experiments previously run.

The foundational work of Stuart et al. [2011] provides an approach based on propensity scores for generalizing the results of experimental interventions to target populations. Our work will leverage this general framework, but introduce methods for optimizing an experimental design to ensure effective generalization performance of resulting estimates. Hartman et al. [2015] similarly uses a combination of matching and weighting to generalize experimental results in-sample to a population average treatment effect on the treated. Other work has also considered weighting-based approaches to generalization [Buchanan et al., 2018].

Dehejia et al. [2019] shows how to use an outcome-modeling approach to extrapolate effects estimated in one population to a population. In contrast to Hartman et al. [2015] and Stuart et al. [2011], this approach relies on modeling the outcomes and then predicting effects in different locations rather than simply reweighting data observed in-sample.

Dahabreh et al. [2018] provides a variety of estimation methods to generalize to a target population, including doubly-robust methods. Rudolph and van der Laan [2017], likewise, provides a doubly-robust targeted maximum likelihood estimator for transporting effects.

There has also been work focused particularly on identification in this setting. Dahabreh et al. [2019] defines a rigorous sampling framework for describing generalizability of experimental results and identifiability conditions through the g-formula. Pearl and Bareinboim [2011] lays out a general framework for determining identifiability of effects generalized to new populations through.

Miratrix et al. [2018] and Coppock et al. [2018] challenge the premise of the necessity for generalization due to the rarity of heterogeneous treatment effects. These studies specifically focused on survey experiments, however, and it isn't truly up for debate that many important objects of study have important heterogeneous components [Allcott, 2015, Vivalt, 2015, Dehejia et al., 2019].

Experimental design.

The standard practice for experimental design is blocking [Greevy et al., 2004], in which units are divided into clusters and then a fixed number of units within each cluster are assigned to treatment. This ensures balance on the cluster indicators within the sample. Higgins et al. [2016] provides a blocking scheme based on k-nearest-neighbors that can be calculated more efficiently than the "optimal" blocking of [Greevy et al., 2004].

Kallus [2018] takes an optimization approach to the problem of experimental design. This work optimizes treatment allocations based on in-sample measures of balance (particularly with respect to kernel means), showing how assumptions of smoothness are necessary to improve on simple Bernoulli randomization.

Rerandomization approaches simply draw allocations randomly until one is located which meets the pre-specified balance criteria. This is also the basis of our proposed method. Morgan et al. [2012] analyzes the rerandomization procedure of discarding randomized assignments that have more in-sample imbalance than a pre-specified criteria in terms of Mahalanobis distance. Li et al. [2018] provides asymptotic results for rerandomization that does not rely on distributional assumptions on the covariates.

Harshaw et al. [2019] provides an efficient method for obtaining linear balance using a Gram-Schmidt walk. Their algorithm includes a robustness-balance tradeoff tuneable by a parameter in their algorithm, and provides useful tools for analyzing experimental design which we use in our theoretical analyses in Section 5.

All aforementioned work on experimental design places as its objective estimation of effects on the sample (i.e.

it optimizes for the sample average treatment effect). This work departs by considering the alternative objective of prioritizing estimation on a target population (i.e. the population average treatment effect).

5 Analysis

In this section we will analyze the expectation and variance of our importance-weighted estimator in Eq. 2 with Target Balance in Definition 1.

Section 5.1 shows that using the importance-weighted estimator with Target Balance results in an unbiased estimator of the target's ATE (Theorem 1).

In Section 5.2 we analyze the variance. In Section 5.2.1, Corollary 1 shows that when the dimension of the covariates $d = 1$, for a finite sample size n, Target Balance reduces the variance. Moreover, among all reasonable balance criteria with rejection probability at most α (including Source Balance), Target Balance achieves the optimal variance reduction (Theorem 2). Section 5.2.2 shows that when $d \geq 1$, when the sample size is large, Target Balance reduces the variance (Theorem 3) and achieves a lower variance than Source Balance (Theorem 4).

5.1 Expectation

In this section we will show that our importance-weighted estimator in Eq. 2 is an unbiased estimator of the target's ATE with Target Balance:

Theorem 1. *Let $\hat{\tau}_Y^T$ be the importance-weighted estimator in Equation 2. When $n_0 = n_1 = n/2$: $\mathbb{E}^S_{\mathbf{X},\mathbf{Y},\mathbf{Z}_{\phi_T}}\left[\hat{\tau}_Y^T\right] = \tau_Y^T$.*

The proof makes use of the fact that the conditional distributions of Y given X in both the source and the target are the same ($p_S(Y|X) = p_T(Y|X)$), and therefore $\frac{p_T(X)}{p_S(X)} = \frac{p_T(X,Y)}{p_S(X,Y)}$.

5.2 Variance

In this section we analyze the variance. We use $\tilde{Y}_i^a$ and $\tilde{y}_i^a$ to denote $W_i Y_i^a$ and $w_i y_i^a$ for $a \in \{0, 1\}$.

5.2.1 Finite Sample Size Variance Reduction for $d = 1$

In this section we will show that when X is a 1-dimensional random variable and the sample size is finite, Target Balance reduces the variance compared to complete randomization. Moreover, among all symmetric balance conditions (defined below) with rejection probability at most α (including Source Balance), Target Balance achieves the optimal variance reduction. The variance can be decomposed into 2 terms (Lemma 3) where the second term does not depend on the balance. The first term is the variance of a 1d symmetric random variable, and Target Balance corresponds to truncating the tail, which results in the largest variance reduction (Theorem 2).

Let $\rho(\mathbf{x}, \mathbf{Z}) \in \{0, 1\}$ denote a function that depends on only $\mathbf{x}$ and $\mathbf{Z}$, and satisfies the symmetric condition $\rho(\mathbf{x}, \mathbf{Z}) = \rho(\mathbf{x}, -\mathbf{Z})$. This definition captures all reasonable balance conditions (including Source Balance) where $\rho = 1$ denotes acceptance and $\rho = 0$ denotes rejection. Note that the constant function $\rho(\mathbf{x}, \mathbf{Z}) = 1$ for all $\mathbf{x}, \mathbf{Z}$ also satisfies the criteria $\rho(\mathbf{x}, \mathbf{Z}) = \rho(\mathbf{x}, -\mathbf{Z})$, and $\rho = 1$ becomes the entire sample space. We proceed to compare Target Balance with any ρ satisfying the criteria above.

First we note that by the law of total variance:

Lemma 1. *For any function $\rho(\mathbf{x}, \mathbf{Z}) \in \{0, 1\}$ satisfying $\rho(\mathbf{x}, \mathbf{Z}) = \rho(\mathbf{x}, -\mathbf{Z})$:*

$$\begin{aligned} var^S_{\mathbf{X},\mathbf{Y},\mathbf{Z}_\rho}(\hat{\tau}_Y^T) =& \mathbb{E}^S_{\mathbf{X}}\left[var^S_{\mathbf{Y},\mathbf{Z}_\rho}(\hat{\tau}_Y^T|\mathbf{X})\right] \\ &+ var^S_{\mathbf{X}}(\frac{1}{n}\sum_{i=1}^n W_i(\beta_1 - \beta_0)^T X_i) \ . \end{aligned}$$

Note that the second term does not depend on ρ. Therefore we focus on analyzing the variance conditioned on $\mathbf{X} = \mathbf{x}$ in this section, and the result for $\text{var}^S_{\mathbf{X},\mathbf{Y},\mathbf{Z}_\rho}(\hat{\tau}_Y^T)$ easily follows from $\text{var}^S_{\mathbf{Y},\mathbf{Z}_\rho}(\hat{\tau}_Y^T|\mathbf{x})$.

Let $C_i = \frac{Y_i^1 + Y_i^0}{2}$, $c_i = \frac{y_i^1 + y_i^0}{2}$, $\beta = \frac{\beta_1 + \beta_0}{2}$, $\mathcal{E} = \frac{\mathcal{E}_1 + \mathcal{E}_0}{2}$ and $\sigma^2_{\mathcal{E}} = \text{var}(\mathcal{E})$. The variance of the importance weighted estimator can be written as

Lemma 2. *Let $n_0 = n_1 = n/2$. For any function $\rho(\mathbf{x}, \mathbf{Z}) \in \{0, 1\}$ satisfying $\rho(\mathbf{x}, \mathbf{Z}) = \rho(\mathbf{x}, -\mathbf{Z})$:*

$$var_{\mathbf{Z}_\rho}(\hat{\tau}_Y^T|\mathbf{x}, \mathbf{y}) = \frac{4}{n^2}\mathbb{E}_{\mathbf{Z}_\rho}\left[\left(\sum_{i=1}^n Z_i w_i c_i\right)^2 \middle| \mathbf{x}, \mathbf{y}\right]$$

Using the law of total variance and the fact that $W_i C_i = W_i X_i \beta + W_i \mathcal{E}$ and $\mathbb{E}[\mathcal{E}|\mathbf{x}] = 0$ we have:

Lemma 3. *Let $n_0 = n_1 = n/2$. For any function $\rho(\mathbf{x}, \mathbf{Z}) \in \{0, 1\}$ satisfying $\rho(\mathbf{x}, \mathbf{Z}) = \rho(\mathbf{x}, -\mathbf{Z})$:*

$$\begin{aligned} &var^S_{\mathbf{Y},\mathbf{Z}_\rho}(\hat{\tau}_Y^T|\mathbf{x}) \\ &= \frac{4}{n^2}\beta^2 \mathbb{E}_{\mathbf{Z}_\rho}\left[\left(\sum_{i=1}^n w_i x_i Z_i\right)^2 \middle| \mathbf{x}\right] + \frac{6}{n^2}\sigma^2_{\mathcal{E}}\sum_{i=1}^n w_i^2. \end{aligned}$$

We note that the design affects only the first term in the above decomposition. Let $V := \frac{2}{n}\sum_i Z_i w_i x_i = \frac{2}{n}\tilde{\mathbf{x}}^T\mathbf{Z}$ and let $B := V\text{var}(V)^{-1/2}$. Recall that the Malahanobis distance $M(\frac{2}{n}(\mathbf{wx})^T\mathbf{Z}) = ||B||^2$. Re-randomization procedure corresponds to truncating B

where B is a mean zero random variable (as Z_i's are random variables) that is symmetric about zero.

It is easy to show that the best way to truncate a symmetric random variable B to minimize the variance is to truncate the tail symmetrically $\|B\|^2 < a$ for some threshold a. Therefore Target Balance reduces the variance, and among all the balance conditions with rejection probability at most α (including Source Balance), Target Balance achieves the optimal variance reduction.

Theorem 2. *Let $n_0 = n_1 = n/2$ and $d = 1$. Let $\rho(\mathbf{x}, \mathbf{Z})$ be a function satisfying $\rho(\mathbf{x}, \mathbf{Z}) = \rho(\mathbf{x}, -\mathbf{Z})$ and $\mathbb{P}(\rho = 1|\mathbf{x}) \geq 1 - \alpha$. Then:*

$$var^S_{\mathbf{Y},\mathbf{Z}_{\phi^\alpha_T}}(\hat{\tau}^T_Y|\mathbf{x},) \leq var^S_{\mathbf{Y},\mathbf{Z}_\rho}(\hat{\tau}^T_Y|\mathbf{x}).$$

Applying Theorem 2 with ρ being the constant function $\rho(\mathbf{x}, \mathbf{Z}) = 1$ for all $\mathbf{x}, \mathbf{Z}$, we have:

Corollary 1. *When $d = 1$ and $n_0 = n_1 = n/2$, using Target Balance reduces the variance compared to complete randomization:*

$$var^S_{\mathbf{Y},\mathbf{Z}_{\phi_T}}(\hat{\tau}^T_Y|\mathbf{x}) \leq var^S_{\mathbf{Y},\mathbf{Z}}(\hat{\tau}^T_Y|\mathbf{x})$$

5.2.2 Asymptotic Variance Reduction for $d \geq 1$

In this section we show that when the sample size is large, Target Balance reduces the variance and achieves a lower variance than Source Balance. We discuss the case of finite sample size in the appendix.

From [Li et al., 2018], the importance weighted estimator can be decomposed into 2 components: part 1 is related to the covariates and part 2 is unrelated. Only part 1 is reduced by rerandomization while part 2 is unaffected. The covariates can be chosen to be the importance-weighted covariates (Target Balance) or the unweighted covariates (Source Balance). Since the importance-weighted covariates aligns better with the importance-weighted outcomes, part 1 will be larger and therefore the reduction by re-randomization will be larger.

In this section we condition on $\mathbf{x}$ and $\mathbf{y}$ so the randomness only comes from $\mathbf{Z}$. Similar to Section 5.2.1, first we note that by the law of total variance:

Lemma 4. *For any function $\rho(\mathbf{x}, \mathbf{Z}) \in \{0, 1\}$ satisfying $\rho(\mathbf{x}, \mathbf{Z}) = \rho(\mathbf{x}, -\mathbf{Z})$:*

$$\begin{aligned} var^S_{\mathbf{X},\mathbf{Y},\mathbf{Z}_\rho}(\hat{\tau}^T_Y) =& \mathbb{E}^S_{\mathbf{X},\mathbf{Y}} var_{\mathbf{Z}_\rho}(\hat{\tau}^T_Y|\mathbf{X},\mathbf{Y}) \\ &+ var^S_{\mathbf{X},\mathbf{Y}}(\sum_{i=1}^n W_i(Y_i^1 - Y_i^0)) \end{aligned}$$

Since the second term does not depend on ρ, we focus on analyzing the variance conditioned on $\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}$ in this subsection. The result for $\text{var}^S_{\mathbf{X},\mathbf{Y},\mathbf{Z}_\rho}(\hat{\tau}^T_Y)$ easily follows from $\text{var}_{\mathbf{Z}_\rho}(\hat{\tau}^T_Y|\mathbf{X},\mathbf{Y})$.

Conditioning on $\mathbf{x}$ and $\mathbf{y}$, Li et al. [2018] state that if the following conditions (Condition 1 in [Li et al., 2018]) are satisfied, finite central limit theorem implies that $(\hat{\tau}^T_Y, \frac{2}{n}\tilde{\mathbf{x}}^T\mathbf{Z})$ approaches a normal distribution as n goes to infinity. Let $\text{avg}(\tilde{\mathbf{y}})$ and $\text{avg}(\tilde{\mathbf{x}})$ denote the average of the rows of $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$. As $n \to \infty$:

- The finite population variances and covariance $cov(\tilde{\mathbf{x}}), cov(\tilde{\mathbf{y}}^1), cov(\tilde{\mathbf{y}}^0), cov(\tilde{\mathbf{y}}^1 - \tilde{\mathbf{y}}^0), cov(\tilde{\mathbf{y}}^1, \tilde{\mathbf{x}})$ and $cov(\tilde{\mathbf{y}}^0, \tilde{\mathbf{x}})$ have limiting values.
- $\max_{1\leq i\leq n} |\tilde{y}^a_i - \text{avg}(\tilde{\mathbf{y}})^a|^2/n \to 0$ for $a \in \{0, 1\}$ and $\max_{1\leq i\leq n} \|\tilde{x}_i - \text{avg}(\tilde{\mathbf{x}})\|_2^2/n \to 0$

We apply Corollary 2 in Li et al. [2018] to give the expression for the asymptotic variance of $\hat{\tau}^T_Y$ under Mahalanobis balance condition. Let as-var denote the variance of the asymptotic sampling distribution of a sequence of random variables. Applying Corollary 2 in [Li et al., 2018] to our case with covariates $\tilde{\mathbf{x}}$ and $\mathbf{x}$ and the weighted outcome $\tilde{\mathbf{y}}$ directly yields the following result showing both Target Balance and Source Balance reduce the variance.

Theorem 3 (Corollary 2 in [Li et al., 2018])**.** *When $n_0 = n_1 = n/2$:*

$$\begin{aligned} &as\text{-}var_{\mathbf{Z}_{\phi_S}}\left(\hat{\tau}^T_Y|\mathbf{x},\mathbf{y}\right) \\ &= \lim_{n\to\infty} var_{\mathbf{Z}}(\hat{\tau}^T_Y|\mathbf{x},\mathbf{y})(1 - (1 - v_{d,a})R^2_{\mathbf{x}}), \\ &as\text{-}var_{\mathbf{Z}_{\phi_T}}\left(\hat{\tau}^T_Y|\mathbf{x},\mathbf{y}\right) \\ &= \lim_{n\to\infty} var_{\mathbf{Z}}(\hat{\tau}^T_Y|\mathbf{x},\mathbf{y})(1 - (1 - v_{d,a})R^2_{\tilde{\mathbf{x}}}), \end{aligned}$$

where $R^2_{\tilde{\mathbf{x}}} = Corr(\hat{\tau}^T_Y, \frac{2}{n}\tilde{\mathbf{x}}^T\mathbf{Z})$, $R^2_{\mathbf{x}} = Corr(\hat{\tau}^T_Y, \frac{2}{n}\mathbf{x}^T\mathbf{Z})$ and $v_{d,a} = \frac{P(\chi^2_{d+2}\leq a)}{P(\chi^2_d\leq a)}$.

We now show that Target Balance has a smaller variance than Source Balance. We use the following equivalent expressions for $R^2_{\mathbf{x}}$ and $R^2_{\tilde{\mathbf{x}}}$. Let $Q = \frac{n}{n-1}\left(\mathbf{I}_d - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)$ where $\mathbf{I}_d$ is an identity matrix of dimension d. Recall that $c_i = \frac{y_i^0 + y_i^1}{2}$. Let $\mathbf{c} := (c_1, \cdots, c_n)$ and $\tilde{\mathbf{c}} = \mathbf{w}\mathbf{c}$. We will show that:

Lemma 5. *When $n_0 = n_1 = n/2$:*

$$R^2_{\mathbf{x}} = \sqrt{\frac{\|Q\tilde{\mathbf{c}}\|^2 - \min_{\hat{\beta}} \|Q\tilde{\mathbf{c}} - Q\mathbf{x}\hat{\beta}\|^2}{\|Q\tilde{\mathbf{c}}\|^2}}.$$

$$R^2_{\tilde{\mathbf{x}}} = \sqrt{\frac{\|Q\tilde{\mathbf{c}}\|^2 - \min_{\hat{\beta}} \|Q\tilde{\mathbf{c}} - Q\tilde{\mathbf{x}}\hat{\beta}\|^2}{\|Q\tilde{\mathbf{c}}\|^2}}.$$

Intuitively $R^2_{\tilde{\mathbf{x}}}$ and $R^2_{\mathbf{x}}$ describe how well $\tilde{\mathbf{c}}$ is described by a linear function of $\tilde{\mathbf{x}}$ and $\mathbf{x}$, respectively. Because of our model, a linear model in terms of $\tilde{\mathbf{x}} = \mathbf{w}\mathbf{x}$ fits $\tilde{\mathbf{c}} = \mathbf{w}\mathbf{c}$ better than a linear model in terms of

$\mathbf{x}$. Therefore, $R^2_{\tilde{\mathbf{x}}}$ will be larger than $R^2_{\mathbf{x}}$ and using ϕ_T will result in a smaller variance than ϕ_S.

Therefore with the same rejection probability α, Target Balance has a lower variance than Source Balance.

Theorem 4. *When $n_0 = n_1 = n/2$, if X_i, Y_i and W_i have finite eighth moment according to the source distribution, with the same rejection probability α:*

$$\textit{as-var}_{\mathbf{Z}_{\phi_T^\alpha}}\left(\hat{\tau}_Y^T|\mathbf{X},\mathbf{Y}\right) \leq \textit{as-var}_{\mathbf{Z}_{\phi_S^\alpha}}\left(\hat{\tau}_Y^T|\mathbf{X},\mathbf{Y}\right)$$

almost surely.

6 Simulations

We perform simulations on the two following models:

Linear Model

$$Y^0 = X + Norm(0,1); Y^1 = 3X + Norm(0,1)$$

Nonlinear Model

$$Y^0 = X^T X + Norm(0,1); Y^1 = 2X^T X + Norm(0,1)$$

We use the following source and target distributions for X. In the source distribution, $X \sim \text{MultivariateNorm}(\mathbf{1}, \mathbf{I})$ where $\mathbf{I}$ is the identity matrix. In the target distribution, $X \sim \text{MultivariateNorm}(\mathbf{1} + \delta, \mathbf{I})$ where δ is a parameter that will be specified later.

We randomly choose an assignment such that $n_1 = n_0 = n/2$. To select the random assignment with the top balance, instead of choosing a fixed threshold α, we select the rejection probability $\alpha = 0.99$ as in Def. 1. To implement this, we draw $100/(1-\alpha)$ assignments at random, calculate their Mahalanobis distance and pick one among the smallest 100 uniformly at random.

If the source and the target distributions are far away, importance weighting can induce large variance. We use the weight clipping technique, in which if the importance weight is larger than a threshold, it will be set to that threshold. It will induce bias but reduce variance, and therefore reduce mean square error (MSE).

We compare 6 methods (WE, CR), (WE, SB), (WE, TB), (UE, CR), (UE, SB) and (UE, TB) by combining the following 2 properties:

Weighted and Unweighted.

- *Weighted Estimator (WE).* We consider the importance weighted estimator in Eq. 2.
- *Unweighted Estimator (UE).* We consider the unweighted estimator which is equivalent to Eq.2 with all weights set to one.

Complete Randomization, Source Balance and Target Balance.

- *Complete Randomization (CR).* This is the randomized assignment without balancing.
- *Source Balance (SB).* This is the rerandomization algorithm seeking Source Balance.
- *Target Balance (TB).* This is the rerandomization algorithm seeking Target Balance as in Definition 1.

We study the MSE of our methods in relation to the 3 following parameters: the sample size n, the importance weights threshold and the distance δ. Recall that in the source distribution, $X \sim \text{MultivariateNorm}(\mathbf{1}, \mathbf{I})$ where $\mathbf{I}$ is the identity matrix and in the target distribution, $X \sim \text{MultivariateNorm}(\mathbf{1} + \delta, \mathbf{I})$.

Sample Size. In this experiment for both models we vary the sample size from 500 to 9500 with step size 500 and set the number of covariates to 10. For the linear model, $\delta = 0.3$. For the nonlinear model, $\delta = 0.2$. δ is chosen to be small enough so that we do not need weight clipping. For each sample size we repeat the experiment 500 times. There is no importance weight threshold. The results are shown and discussed in Figure 1.

Threshold. In this experiment for both models we vary the importance weight threshold from 5, then 10 to 190 with step size 10. We set the number of covariates to 10 and the sample size to be 1000 and $\delta = 0.6$. δ is chosen to be large enough so that weight clipping is necessary. For each threshold we repeat the experiment 500 times. The results are shown and discussed in Figure 3.

Distance δ. In this experiment for both models we vary δ from 0.1 to 0.9 with step size 0.1. We set the number of covariates to 10, the sample size to be 1000 and the importance weight threshold to be 40. From the weight threshold experiment, we know that if the weight threshold is too large, the variance is too high while if the weight threshold is too small, the bias will be too high. Therefore we pick the value 40 as a reasonable weight threshold. For each threshold we repeat the experiment 500 times. The results are shown and discussed in Figure 2. Across all simulations, Target Balance with the Weighted Estimator substantially reduces the MSE.

7 Conclusion

In this work, we've shown that a desire for generalizability should change the way experiments are designed and run. In particular, we argue that balance should be sought on the target population rather than the samples in which randomization will actually be performed. We present a method for designing an experiment along these lines, show theoretically that it is unbiased and more efficient than sample balancing.

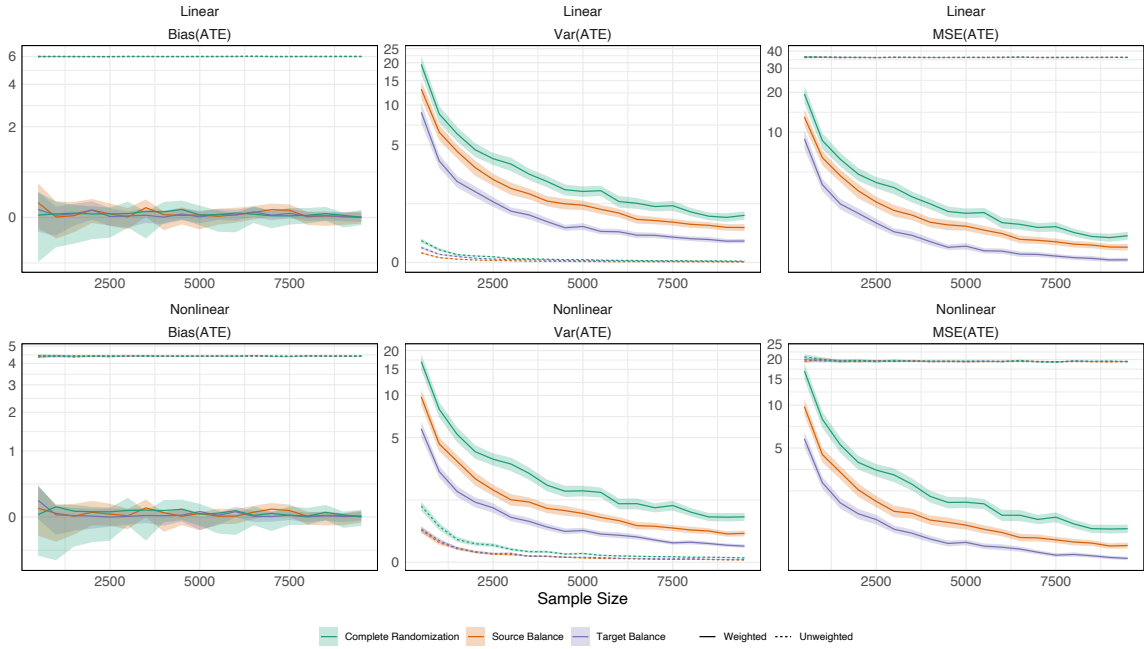

Figure 1: Bias, Variance and MSE as a function of the sample size. All unweighted estimators are biases because they measure the ATE of the source distribution. As there is no importance weight threshold, all weighted estimators are unbiased (Theorem 1) but the weighted estimator with Target Balance has the lowest variance. *The y axes are in log scale.*

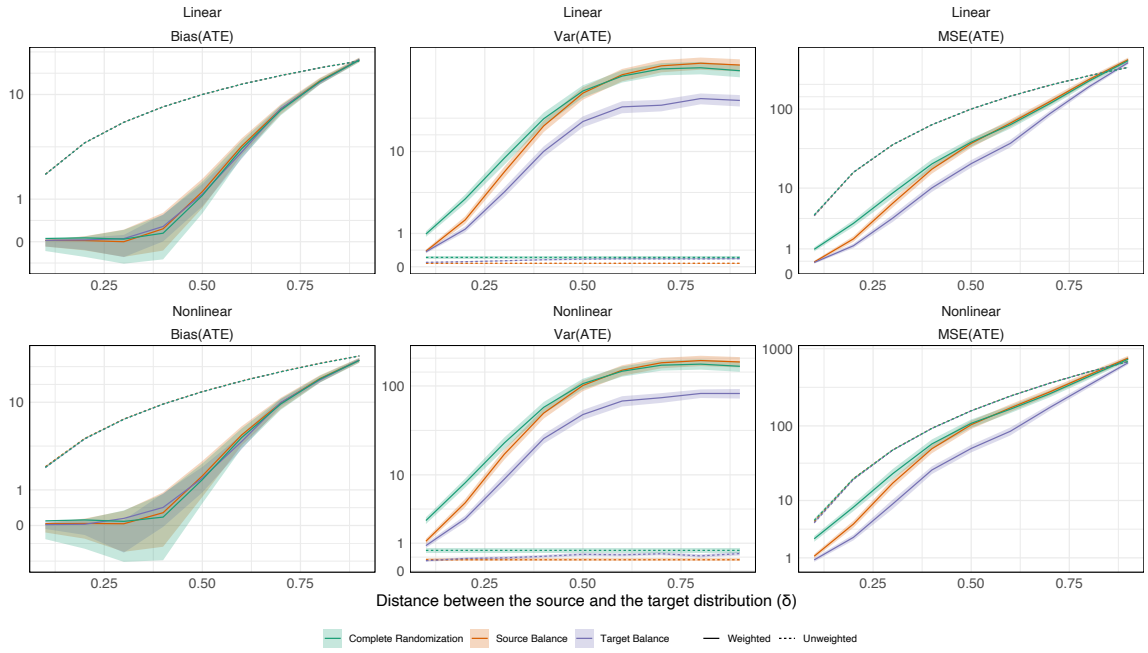

Figure 2: Bias, Variance and MSE as a function of the distance δ (defined in Section 6) between the source and the target distribution. Because of the importance weight threshold, the biases of the importance weighted methods increase as the δ increase. If the distance is too large, the bias of the importance weighted estimators is large, leading to high MSE. However when the distance is not too large, the weighted estimator with Target Balance has the lowest MSE. *The y axes are in log scale.*

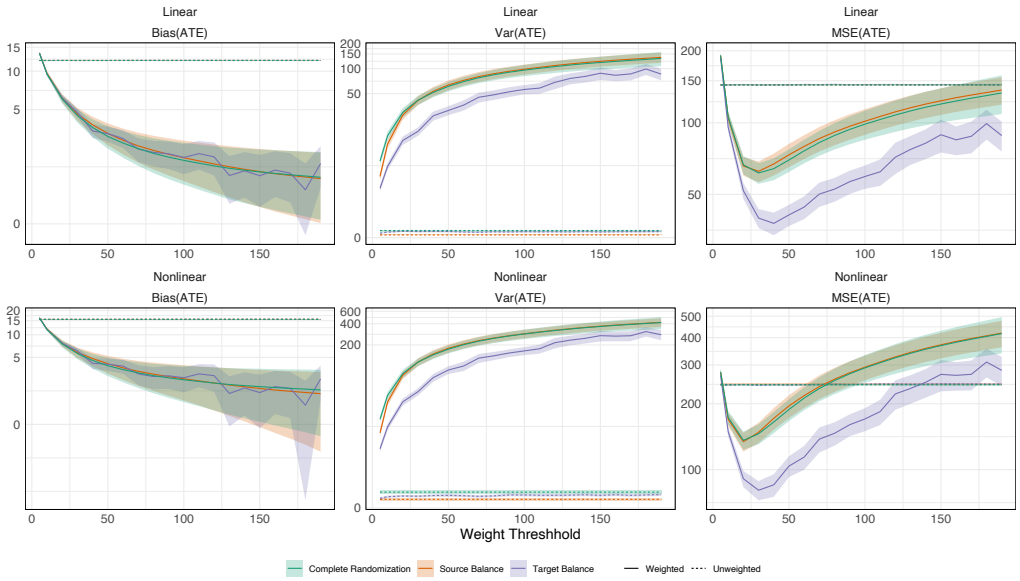

Figure 3: Bias, Variance and MSE as a function of importance weight threshold. As the threshold increases, the bias of the weighted methods decreases and the variance of the weighted methods increases. Therefore there is a threshold when the MSE is minimized. The weighted estimator with Target Balance has the lowest MSE for a reasonably good threshold. *The y axes are in log scale.*

References

Hunt Allcott. Site selection bias in program evaluation. *The Quarterly Journal of Economics*, 130(3):1117–1165, 2015.

Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, Katie R Mollan, Paul E Sax, Eric S Daar, Adaora A Adimora, Joseph J Eron, and Michael J Mugavero. Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society. Series A,(Statistics in Society)*, 181(4):1193, 2018.

Alexander Coppock, Thomas J. Leeper, and Kevin J. Mullinix. Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49):12441–12446, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1808083115. URL `https://www.pnas.org/content/115/49/12441`.

Issa J. Dahabreh, Sarah E. Robertson, Jon A. Steingrimsson, Elizabeth A. Stuart, and Miguel A. Hernan. Extending inferences from a randomized trial to a new target population, 2018.

Issa J. Dahabreh, Sebastien J-P. A. Haneuse, James M. Robins, Sarah E. Robertson, Ashley L. Buchanan, Elisabeth A. Stuart, and Miguel A. Hernán. Study designs for extending causal inferences from a randomized trial to a target population, 2019.

Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii. From local to global: External validity in a fertility natural experiment. *Journal of Business & Economic Statistics*, pages 1–27, 2019.

Eva H. DuGoff, Megan Schuler, and Elizabeth A. Stuart. Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research*, 49 (1):284–303, 2014. doi: 10.1111/1475-6773.12090. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6773.12090`.

Thad Dunning, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, Gareth Nellis, Claire L. Adida, Eric Arias, Clara Bicalho, Taylor C. Boas, Mark T. Buntaine, Simon Chauchard, Anirvan Chowdhury, Jessica Gottlieb, F. Daniel Hidalgo, Marcus Holmlund, Ryan Jablonski, Eric Kramon, Horacio Larreguy, Malte Lierl, John Marshall, Gwyneth McClendon, Marcus A. Melo, Daniel L. Nielson, Paula M. Pickering, Melina R. Platas, Pablo Querubín, Pia Raffler, and Neelanjan Sircar. Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials. *Science Advances*, 5(7), 2019. doi: 10.1126/sciadv.aaw2612. URL `https://advances.sciencemag.org/content/5/7/eaaw2612`.

Jill A Fisher and Corey A Kalbaugh. Challenging assumptions about minority participation in us clini-