Designing Multi-Context Studies for External Validity: Site Selection via Synthetic Purposive Sampling*

Naoki Egami† Diana Da In Lee‡

First Version: July 3, 2023
This Version: August 24, 2023

Abstract

To address the most common and challenging external validity concerns about contexts, an increasingly popular, powerful empirical strategy is a multi-context/multi-site design where researchers conduct causal studies in each site and test whether findings generalize across contexts. Despite its significant potential, there has been little systematic guidance on the fundamental research design question — how should we select study sites for external validity? Existing approaches have well-known challenges: random sampling of sites is often infeasible, while the current practice of purposive sampling is suboptimal without transparent statistical guarantees. In this paper, we propose a general approach, *synthetic purposive sampling* (SPS), which optimally selects diverse study sites for external validity. SPS improves upon conventional purposive sampling by combining ideas from the synthetic control method — it selects diverse sites such that non-selected sites are well approximated by the weighted average of the selected sites. Unlike existing alternatives, SPS can accommodate user-specified, practical constraints. This paper offers a new statistical foundation to design multi-site experimental and observational studies for external validity.

Keywords: External validity, Site Selection, Multi-Site Experiments, Causal inference, Generalization

---

*The proposed methodology is implemented via our open-source software R package, `spsR` (`http://naokiegami.com/spsR`). We thank Elias Naumann, Lukas Stötzer, and Giuseppe Pietrantuono for sharing replication data with us. We appreciate excellent research assistance by Songpo Yang and Xiaolong Yang, and we would like to thank Michael Denly, Michael Findley, Don Green, Jens Hainmueller, Melody Huang, Kosuke Imai, Gary King, Ian Lundberg, John Marshall, Kevin Munger, Cyrus Samii, Fredrik Sävje, Tara Slough, Elizabeth Tipton, Scott Tyson, Chagai Weiss, Anna Wilke, and Teppei Yamamoto, for their thoughtful comments. We also appreciate comments from participants at the Political Methodology Summer meeting, a seminar at Columbia and the workshop, The Methodological Challenges of Meta Analysis in the Social Sciences.

†Assistant Professor, Department of Political Science, Columbia University, New York, NY 10027. Email: naoki.egami@columbia.edu, URL: `https://naokiegami.com`

‡Ph.D. student, Department of Political Science, Columbia University, New York, NY 10027. Email: dl2860@columbia.edu, URL: `https://www.dianadainlee.com`

1 Introduction

Over the last twenty years, social scientists have experienced a credibility revolution and made significant progress toward internal validity, focusing on unbiased estimation of causal effects within a study. Another fundamental, long-standing methodological debate is about *external validity* — how scientists can generalize causal findings beyond a specific study (Shadish *et al.*, 2002). While the question of external validity is multi-dimensional, the essential question to social scientists involves contexts: whether and how can researchers generalize causal findings across different contexts and settings? This is one of the most common and yet most challenging external validity concerns social scientists face in practice.

An increasingly popular, promising strategy to address this question is a multi-context/multi-site experimental and observational study where researchers conduct causal studies in multiple contexts to compare and aggregate findings across contexts.[1] For example, one popular type of multi-site causal study is a multi-country survey experiment that tests how causal findings vary across contexts (e.g., Tomz and Weeks, 2013; Carnes and Lupu, 2016; Valentino *et al.*, 2019; Bassan-Nygate *et al.*, 2023). Such multi-site causal studies are powerful strategies toward external validity because researchers can explicitly exploit across-context heterogeneity rather than extrapolating causal findings from a single context, which often requires untenable assumptions (e.g., Shadish *et al.*, 2002; Gerber and Green, 2012; Tipton, 2013; Allcott, 2015; Dehejia *et al.*, 2019; Blair and McClendon, 2020; Findley *et al.*, 2020; Wilke and Humphreys, 2020; Miratrix *et al.*, 2021; Egami and Hartman, 2022; Slough and Tyson, 2022a; Wilke and Samii, 2023).

Understanding the importance of external validity, an increasing number of scholars deploy such multi-site causal studies. In the top 10 political science journals, the number of multi-site causal studies has increased gradually over time (see Figure 1). There were only a few multi-site studies before 2010, but the number increased steadily since then. Between 2000 and 2022, we find 130 multi-site studies (103 experimental and 27 observational studies). This increasing trend is expected to continue because running experiments in multiple contexts has become easier and cheaper (e.g., survey companies offer online panels in many countries with low cost) and observational identification strategies have been widely used. Multi-site causal studies have also been strongly supported by initiatives like Metaketa by EGAP (e.g., Dunning *et al.*, 2019) and education and microcredit experiments by JPAL (Banerjee *et al.*, 2015a,b).

[1]We define a multi-site causal study to be a study where researchers have (experimental or observational) identification strategies for internal validity *within* each site and researchers compare results across sites for external validity.

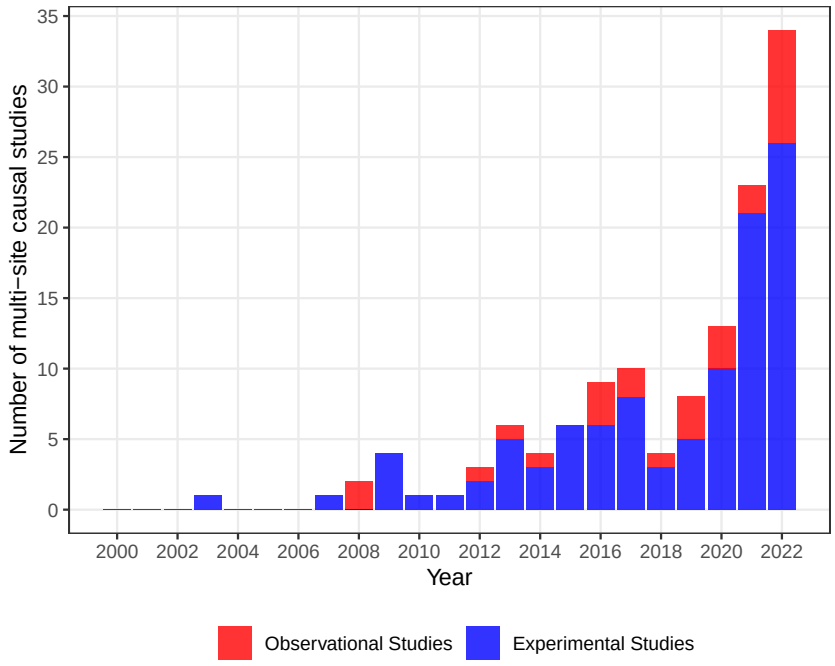


Figure 1: **Increasingly Popular Multi-Site Causal Studies.**
Note: Blue (red) bars represent multi-site experimental (observational) studies. The plot is based on a review of articles published in the top 10 political science journals from 2000 to 2022. See Appendix A.1 for information about how the review was conducted.

Despite this significant and promising increase in multi-site causal studies, there has been little systematic guidance on the fundamental research design question — how should we select study sites for external validity? Unless sites are selected systematically, results from selected sites are not generalizable to broader contexts and cannot improve external validity credibly. Unfortunately, existing strategies available to applied social scientists are limited and have well-known challenges.[2] Broadly speaking, there are two classes of existing strategies: random sampling and purposive sampling.

[2]This paper focuses on multi-site causal studies in political science and related social science fields where the number of study sites is small. Our literature review of the top 10 political science journals finds that the median number of study sites is 3 and the 80th percentile is 6.6. Toward the end of the introduction, we also discuss the important literature about multi-site causal studies in education and health research (e.g., Raudenbush and Liu, 2000; Tipton, 2013), where the number of study sites is much larger and other sampling strategies have been feasible.

First, if feasible, random sampling of sites is the most powerful approach to make generalizable causal claims from multi-site studies. As random assignment of treatments is the gold standard for internal validity, random sampling of sites is the gold standard for external validity. However, unfortunately, in realistic settings of social science studies, random sampling from all the sites of theoretical interest is often infeasible because of logistical and ethical reasons. Indeed, our literature review of the top political science journals found only 2 studies that use random sampling (less than 2 % of all multi-site experiments we reviewed).

Given the difficulty of random sampling, researchers often rely on purposive sampling (Shadish *et al.*, 2002), which is a non-probability sampling technique that selects sites with "theoretical purposes." While it has a number of well-developed variants in the literature, in the practice of empirical studies, the most popular version of purposive sampling is to select diverse sites such that study sites cover heterogeneous contextual factors (more than 80% of multi-site experiments justify their site selection in this manner). For example, when studying attitudes toward immigrants using survey experiments (e.g., Naumann *et al.*, 2018; Valentino *et al.*, 2019), researchers would select diverse countries with different sizes of immigrant populations, GDP, and unemployment rates.

Although the current practice of purposive sampling has some methodological benefits, it suffers from several key challenges. First, because researchers currently select diverse sites mostly by hand, they are often forced to focus on one or two site-level variables, even if other potentially relevant factors exist (our literature review finds that the average number of covariates researchers diversify is 2.11). Second, the process of purposive sampling is often not transparent or not reproducible (Fearon and Laitin, 2008). Finally, purposive sampling is often not formally connected to subsequent statistical analyses, and, as a result, the current practice has no explicit statistical guarantees. Overall, in the words of Olsen *et al.* (2013), the current practice can be seen as "stratified convenience sampling," i.e., researchers carefully discuss one or two contextual factors to stratify, but they choose the most convenient sites after stratification.

In this paper, we develop a novel approach to optimally select study sites for external validity. Our goal is to keep various benefits of purposive sampling, such as practicality and interpretability, while providing transparency and a statistical foundation. In particular, we propose *synthetic purposive sampling* (SPS), which improves upon conventional purposive sampling by combining ideas from the synthetic control method (Abadie *et al.*, 2010). SPS selects diverse sites such that non-selected sites can be well approximated by the weighted average of the selected sites. By doing so, even without random sampling, we can make the weighted average of selected sites representative of all the sites, including non-selected sites.

	Statistical Foundation	Transparency	Practical Feasibility	Incorporate Domain Knowledge
Random Sampling	Gold Standard	High	Low	Difficult
Purposive Sampling (the current practice)	Unclear	Low	High	Easy
Synthetic Purposive Sampling	Connected to SCM Optimally Diversify	High	High	Easy

Table 1: **Comparison of Existing Alternatives and Synthetic Purposive Sampling.** *Note:* SPS improves upon conventional purposive sampling by combining ideas from the synthetic control method (SCM). SPS has both practicality and statistical foundation, only one of which random and purposive sampling methods have.

The proposed SPS has several desirable properties and overcomes shortcomings of existing methods (see Table 1 for the summary). First, it is a flexible approach, unlike random sampling. SPS can accommodate logistical, theoretical, and ethical constraints, e.g., researchers cannot run a survey experiment in China because a survey firm does not recruit online survey participants there. SPS will select the optimal set of study sites conditional on these user-specified constraints. Second, it is transparent. Using SPS, researchers can clarify all the factors and constraints that have affected site selection. Importantly, SPS can explicitly incorporate many site-level covariates, unlike the current practice of purposive sampling that focuses only on one or two variables. Finally, SPS has a clear statistical foundation, unlike the current practice of purposive sampling. We prove that the SPS estimator minimizes the worst-case mean squared error, within a large class of weighted average estimators that includes conventional meta-analysis estimators. SPS possesses both practicality and statistical foundation, whereas random and purposive sampling methods offer only one of these features.

Overall, the proposed method offers a new statistical foundation to design multi-site experimental and observational studies for external validity. This paper takes a *prospective* approach to explicitly design multi-site causal studies for external validity upfront *before* data collection (see also Shadish *et al.*, 2002; Tipton, 2013; Chassang and Kapon, 2022; Slough and Tyson, 2022b). While it is currently common to think about external validity only at the final stage of studies *after* data collection (e.g., start to worry about external validity when writing up papers), such post hoc adjustment requires strong and often untenable assumptions, especially when external validity concerns are about contexts. The proposed SPS allows researchers to explicitly address external validity concerns about contexts upfront through their research design.

To provide concrete ideas about potential applications, Section 2 introduces several examples

of multi-site experimental and observational studies. In Section 3, we review existing strategies and clarify their methodological challenges. We then introduce SPS in Section 4 and discuss how to aggregate causal evidence from multiple sites in Section 5. In Section 6, we discuss how to use SPS step by step and also clarify potential limitations. All of our methods can be implemented via the forthcoming companion R package `spsR`. In Section 7, we provide an empirical application based on a multi-site survey experiment by Naumann *et al.* (2018). In Section 8, we discuss implications and connections to other important large literature, such as case selection in qualitative case studies (e.g., Lieberman, 2005; Fearon and Laitin, 2008; Seawright and Gerring, 2008; Herron and Quinn, 2016).

This paper makes three contributions. First, we contribute to the growing literature on external validity (e.g., Shadish *et al.*, 2002; Tipton, 2013; Bareinboim and Pearl, 2016; Munger, 2019; Blair and McClendon, 2020; Findley *et al.*, 2020; Vivalt, 2020; Miratrix *et al.*, 2021; Egami and Hartman, 2022; Slough and Tyson, 2022a,b; Wilke and Samii, 2023). In particular, this paper examines how to design multi-site studies that are increasingly popular as a strategy to address external validity concerns about contexts. We propose SPS as a general strategy to select diverse sites for external validity, while taking into account various practical constraints researchers face. While we focus on the question about contexts in this paper, other dimensions of external validity, such as populations, outcomes, and treatments, are also essential. We refer to relevant papers throughout this paper and provide more discussions in Section 8.2.

Second, this paper also builds on the large methodological literature on multi-site studies (e.g., Raudenbush and Liu, 2000; Tipton, 2013; Tipton *et al.*, 2014; Tipton and Peck, 2017). For example, Tipton (2013) and Tipton *et al.* (2014) developed a stratified sampling approach by combining ideas of balanced sampling and cluster analysis. These methods are designed for and successful in education and health research where the number of study sites is relatively large. For example, these papers consider settings where "the sample would typically include between 20 and 60 schools or districts" (p.112; Tipton, 2013). In contrast, our paper focuses on settings common in political science and related social science fields where the number of study sites is small (while the sample size in each site is relatively large). Indeed, our literature review of the top political science journals finds that the median number of study sites is 3 and the 80th percentile is 6.6. Our proposed SPS approach is specifically designed for this small sample regime by combining ideas from conventional purposive sampling and the synthetic control method, which was also developed for the small sample regime.

Finally, our paper builds on the literature on the synthetic control method (Abadie *et al.*, 2010). Methodologically, our optimization problem is similar to that of the recent synthetic

design (e.g., Abadie and Zhao, 2021; Doudchenko *et al.*, 2021) that combines the synthetic control method and experimental design to choose treatment assignment for internal validity. The main difference is that SPS selects sites for external validity (rather than treatments for internal validity), which lead to different causal estimands, constraints we add to the main optimization problem, and estimators. We introduce them step by step in Sections 4 and 5.

2 Examples of Multi-Site Causal Studies

As reviewed in the introduction, the number of multi-site causal studies has gradually increased over time. To provide concrete ideas about potential applications of our method, this section provides several illustrative examples of multi-site causal studies.

2.1 Survey Experiments

A survey experiment is the most popular type of multi-site experiment in recent years. Most of these multi-site survey experiments are implemented at the country-level (89% of the multi-site survey experiments in our literature review were multi-country experiments). Such multi-country survey experiments are likely to keep increasing as popular online platforms like YouGov and Lucid can recruit survey respondents across the world.[3]

For example, Lupu and Wallace (2019) study the conditions under which people approve or disapprove of human rights abuses by their governments, using a series of survey experiments. Because recent experiments on the effects of international institutions have disproportionately focused on the United States, they conducted experiments in India, Israel, and Argentina, which have "different histories of state repression and differing experiences with international institutions" (p.412). Magni and Reynolds (2021) examine voter preferences about LGBT candidates using conjoint experiments in the United States, United Kingdom, and New Zealand, which all have single-member district election systems but have "varying degrees of LGBT representation, differing levels of legal progress and resistance to LGBT rights, and different attitudes of parties toward LGBT rights" (p.1200).

Some studies select a larger number of study sites. For example, a prominent study by Valentino *et al.* (2019) examines economic and cultural determinants of immigrant support in 11 countries. They explicitly justify their site selection by diversifying key contextual factors. "Our studies span a wide range of developed democracies. ... The broadly cross-national quality

[3]Lucid (part of Cint now) offers surveys in more than 130 countries (`https://www.cint.com/consumer-insights-exchange`), YouGov covers more than 70 markets (`https://business.yougov.com/product/realtime/international-omnibus`), and Dynata covers more than 45 countries (`https://www.dynata.com/market-researcher-solutions/global-audiences/`).

of the current collection is perhaps its strongest feature, since it allows us to explore the degree to which a specific explanation generalizes across a wide variety of countries with different histories, governmental institutions, economies, cultures and immigration patterns" (Valentino *et al.*, 2019, p.1207). See Table OA-1 in Appendix A.1.1 for more examples of multi-site survey experiments.

2.2 Field Experiments

A multi-site field experiment is one of the most powerful strategies to address external validity. In political science, Metaketa initiative by EGAP is the most famous collective effort using multi-site field experiments. By the end of 2022, four Metaketa projects have conducted multi-country experiments to study various topics, including natural resource governance (Slough *et al.*, 2021) and community policing (Blair *et al.*, 2021), and each project has included 6 countries. As in many field experiments, the site selection process was heavily constrained by various logistical and ethical constraints, such as the availability of high-quality local partners and the feasibility of conducting randomized experiments, and thus, random sampling of sites was infeasible. Other large-scale collective efforts include graduation programs and microcredit experiments by JPAL (Banerjee *et al.*, 2015a,b).

Multi-site field experiments have been conducted not only by the aforementioned large-scale initiatives but also by individual research teams. For example, an influential study by Green *et al.* (2003) conducted door-to-door canvassing experiments in six different cities in the US. "Although the canvassing sites cannot be construed as a random sample of municipal elections occurring nationwide, our study is strengthened by the fact that the get-out-the-vote campaigns took place in very different political and demographic settings. Some elections were tightly contested; others were devoid of meaningful competition. Some sites have large populations of racial and ethnic minorities; others are predominantly white. Our aims in drawing from such a diverse collection of sites are twofold: to better gauge the average treatment effect of canvassing and to examine whether the treatment effects vary systematically with electoral competitiveness or other characteristics of the sites" (p.1086). Other illustrative examples include Valenzuela and Michelson (2016) and Choi *et al.* (2021). Please see Table OA-2 in Appendix A.1.1 for more examples of multi-site field experiments.

2.3 Observational Studies

External validity is equally important in experimental and observational studies (Westreich *et al.*, 2019; Egami and Hartman, 2022). Just as multi-site experiments are often constrained by logistical constraints such as the availability of local partners and online survey panels, multi-

site observational studies are severely constrained by the availability of credible observational identification strategies, and thus, random sampling of sites is rarely feasible.

While the number of multi-site observational studies is smaller than that of multi-site experimental studies (see Figure 1), there are many influential papers. For example, Dehejia *et al.* (2019) studied the impact of fertility on women's labor-force participation across more than 50 countries using a natural experiment idea proposed by Angrist and Evans (1998). Eggers *et al.* (2018) find that scholars deploy regression discontinuity design to measure the effects of population threshold-based policies on political and economic outcomes in many different contexts, covering 12 countries on four continents. Cavaille and Marshall (2019) examine the causal effect of education on anti-immigration attitudes with instrumental variables by exploiting six compulsory schooling reforms in five Western European countries — Denmark, France, Great Britain, the Netherlands, and Sweden. Please see Appendix A.1.2 for more examples of multi-site observational studies.

3 Existing Strategies and Their Methodological Challenges

3.1 Random Sampling

Random sampling of sites is the gold standard method for external validity (Särndal *et al.*, 2003). Its biggest advantage is that randomly selected study sites are representative of a population of sites that researchers are interested in. Thus, researchers are protected from both known and unknown systematic bias in site selection. Without complex adjustment or stringent assumptions, causal findings learned from selected sites are generalizable to a broader population of sites. As random assignment of treatments is the gold standard for internal validity, random sampling of sites is the gold standard to account for site selection bias in external validity analysis.

Despite its theoretical advantage, random sampling is often infeasible in social science applications due to logistical and ethical constraints (see also Findley *et al.*, 2023). For example, scholars might be interested in using survey experiments to study political behavior in Africa, whereas some African countries might not have online survey panels. In American politics, scholars might consider conducting field experiments related to elections in Wisconsin. Yet, it might be ethically and logistically impossible for them to do so, given that it is a battleground state. Indeed, we only find 2 studies that use random sampling of sites in our literature review of multi-site experiments published in the top 10 political science journals.

Another challenge of random sampling is that it might be ineffective when the number of study sites is relatively small, which is the case in political science. Our literature review finds that the median number of sites is 3 and the 80th percentile is 6.6 (see Appendix A.2 for more

details). When the number of sampled sites is small, random sampling can be ineffective because fundamental statistical theorems (e.g., the law of large numbers and the central limit theorem) are not applicable with too small a sample size.

We want to emphasize that researchers should conduct random sampling of sites, if random sampling is logistically and ethically feasible and the number of study sites to be sampled is relatively large. This kind of situation can arise in certain areas, such as in education research settings where scientists often have a relatively large number of schools as sites, while each site has a smaller sample size (e.g., a few hundred) (Olsen *et al.*, 2013).

However, as clarified above, random sampling has been infeasible in most political science applications, and researchers have commonly used an alternative approach of purposive sampling, which we discuss next.

3.2 Conventional Purposive Sampling

Purposive sampling is a class of a non-probability sampling technique that selects sites with "theoretical purposes." It has a long history in the research design literature (Shadish *et al.*, 2002) and has a wide range of well-developed variants, such as typical, extreme, and most similar selection (Seawright and Gerring, 2008).

In practice, the most popular version of purposive sampling is to select diverse sites such that study sites cover a great level of heterogeneity within each contextual factor relevant to a substantive theory of interest. For example, when studying attitudes toward immigrants using survey experiments (e.g., Naumann *et al.*, 2018; Valentino *et al.*, 2019), researchers would select diverse countries with different sizes of immigrant populations, GDP, and unemployment rates (see more examples in our empirical application in Section 7). We find that more than 80% of multi-site experiments justify their site selection by clarifying how selected diverse sites differ in a wide range of contextual factors.

The biggest advantage of purposive sampling is its practicality and interpretability. Unlike strict random sampling, researchers can easily incorporate prior theoretical and domain knowledge as well as logistical and ethical constraints they face. For example, researchers might have strong theoretical and logistical reasons for conducting studies in Uganda — it is a hard test for a given theory, and a researcher has a long-standing local partner that can help her run high-quality experiments. While it is difficult to incorporate these theoretical and logistical considerations in random sampling, purposive sampling can naturally incorporate them.

While purposive sampling has many methodological benefits, its current practice suffers from several key challenges. First, because researchers currently select diverse sites mostly by hand, they are often forced to pick only one or two site-level variables, even when it is likely that other

potentially relevant factors matter (in our literature review, we find that the average number of covariates researchers diversify is 2.11; see Appendix A.2 for more details). Second, the process of purposive sampling is often not transparent or not reproducible (Fearon and Laitin, 2008). Finally, purposive sampling is usually not directly connected to the formal causal inference framework or to subsequent statistical analyses. As a result, the current practice of purposive sampling has no explicit statistical guarantees about external validity analysis. Overall, in the words of Olsen *et al.* (2013), the current practice of purposive sampling can be characterized as "stratified convenience sampling," i.e., researchers carefully discuss one or two contextual factors to stratify, but they choose the most convenient sites after stratification.

4 Synthetic Purposive Sampling

4.1 Overview

In this section, we propose *synthetic purposive sampling* (SPS), which improves upon conventional purposive sampling by combining ideas from the synthetic control method (Abadie *et al.*, 2010). SPS selects diverse sites such that non-selected sites can be well approximated by the weighted average of the selected sites. By doing so, even without random sampling, users can make the weighted averages of selected sites representative.

We will show that SPS naturally introduces diversity in contextual factors, as the current practice of purposive sampling aims to do. Unlike the current practice of purposive sampling, the main benefit of SPS is that it selects diverse sites with transparency and statistical guarantees. Overall, SPS will merge the benefits of random sampling (e.g., statistical guarantees and transparency) and those of purposive sampling (e.g., practicality and interpretability), while accommodating logistical and ethical constraints researchers face in practice.

In Section 4.2, we introduce a framework for site selection. Section 4.3 proposes SPS. After illustrating SPS visually (Section 4.3.1), we describe the optimization problem behind SPS (Section 4.3.2) and explain how users can incorporate various practical constraints (Section 4.3.3).

4.2 Framework for Site Selection

To begin with, we define N potential sites of interest as the target population of sites, which is the target against which external validity of a given substantive theory is evaluated. Specifying the population of sites is equivalent to clarifying the studies' scope conditions, and thus, this choice should be guided by substantive research questions and underlying theories of interest (Findley *et al.*, 2020; Egami and Hartman, 2022).

Among N potential sites of interest, researchers can select N_S sites to run randomized experiments where $N_S \leq N$. To focus on issues of external validity, we consider randomized

experiments here, but it is straightforward to extend our method for observational studies. For example, in our example in Section 7, we define $N = 15$ European countries as the target population of sites, and we select $N_S = 6$ countries as study sites. We assume that researchers use treatment and outcome variables to capture the same underlying theoretical concepts in each site (see Slough and Tyson (2022a) and Wilke and Samii (2023) for the formal discussions and the importance of this assumption).

We now define quantities of interest. For each site $k \in \{1, \ldots, N\}$, we use θ_k to denote the *Site-Specific Average Treatment Effect (ATE)*, which is the average effect of the treatment at site k. For sites researchers select for randomized experiments, they can easily obtain unbiased estimates $\widehat{\theta_k}$, using simple estimators like difference-in-means.

The main issue of external validity is that researchers are not only interested in these estimates in selected sites but also in whether causal conclusions are generalizable to a broader population of N sites. We can define the *Average-Site ATE* as

$$\theta_{AS} \coloneqq \frac{1}{N} \sum_{k=1}^{N} \theta_k, \tag{1}$$

which represents the average of the ATEs across all the N sites of interest. This average-site ATE allows us to investigate whether causal findings in selected sites generalize to a population of N sites, specified by the scope condition. This quantity of interest is widely used and is similar to the common estimand in meta-analyses of multi-site experiments (Gerber and Green, 2012; Dunning *et al.*, 2019; Miratrix *et al.*, 2021).

If researchers can randomly sample study sites, it is straightforward to unbiasedly estimate θ_{AS} using the simple average of difference-in-means in each selected site. However, as discussed in Section 3, random sampling of sites is often infeasible in most social science applications. In the next subsection, we will propose a new approach that selects diverse study sites in order to credibly estimate the average-site ATE.

Note that, even if researchers are not specifically interested in estimating the average-site ATE, our proposed method can also be used as a way to select diverse sites with statistical transparency and flexibility. We discuss this agnostic view of the proposed method in Section 6.

4.3 The Proposed Methodology

We now propose *synthetic purposive sampling* (SPS), which combines two classical ideas from the synthetic control method (Abadie *et al.*, 2010) and purposive sampling (Shadish *et al.*, 2002).

The key methodological idea behind SPS is simple. Like purposive sampling, SPS will select diverse sites. But unlike the current practice of purposive sampling, we will design site selection by explicitly taking into account downstream analyses, i.e., how to use selected sites for

generalization after data collection. In particular, we will use weighted average estimators like the synthetic control method — we will use the weighted average of selected sites to approximate non-selected sites. The weighted average estimator is desirable in several ways. First, it is a safe and conservative estimator as it focuses on interpolation and avoids extrapolation (King and Zeng, 2006). Second, it is also a stable estimator that works well with small sample sizes, and this is critical because the number of study sites N_S is often small (recall that our literature review finds that the median number of experimental sites is 3 and the 80th percentile is 6.6). Finally, it is a familiar estimator to social scientists as most meta-analysis estimators are also weighted average estimators (Miratrix *et al.*, 2021), even though how we construct weights is distinct from conventional meta-analysis estimators.

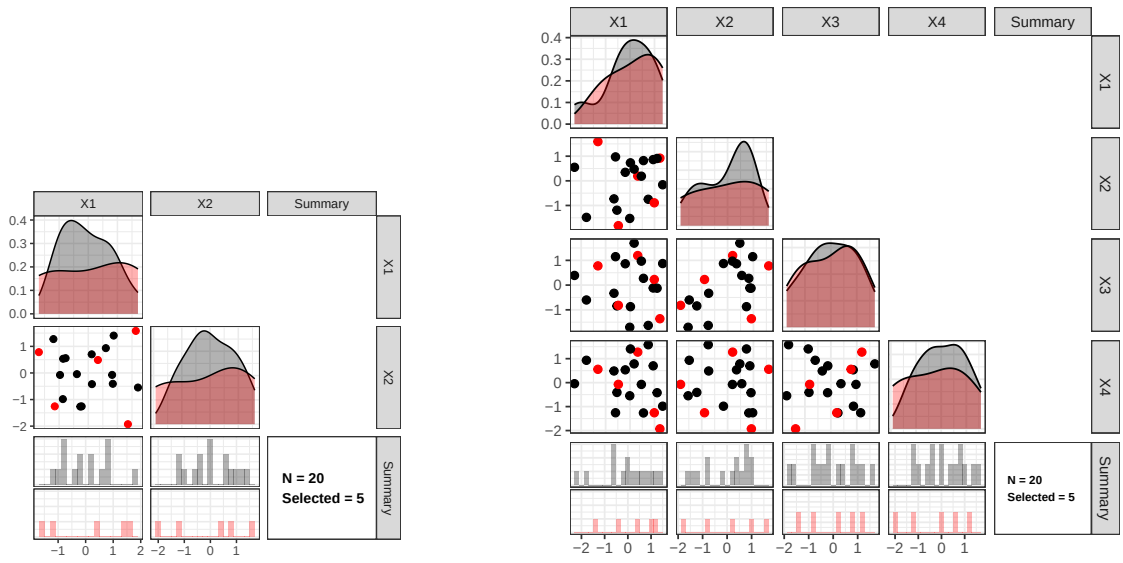
By combining these ideas, SPS will select diverse sites such that non-selected sites can be well approximated by the weighted average of the selected sites. By doing so, even without random sampling, we can make the weighted average of selected sites representative of a population of N sites, including non-selected sites.

More concretely, the first step of SPS is to choose site-level variables $\mathbf{X}_k = (X_{k1}, X_{k2}, \ldots, X_{kL})$ that users want to diversify across sites where L is the number of site-level variables. For example, to run multi-country survey experiments on attitudes toward immigrants, researchers might diversify the size of immigrant populations, GDP, and unemployment rates. Note that, if researchers want to incorporate individual-level covariates, such as the gender and age of respondents within each site, they can incorporate the site-level mean or variance of gender and age as site-level covariates. Formally, we choose contextual moderators that predict differences in the ATEs across sites. Then, SPS will select diverse sites such that variables $\mathbf{X}_k$ of non-selected site $k \in \mathcal{R}$ is well approximated by the weighted average of variables $\mathbf{X}_j$ of selected sites $j \in \mathcal{S}$. Here, $\mathcal{S}$ and $\mathcal{R}$ represent sets of selected and non-selected sites, respectively. For example, we select diverse sites such that the GDP of each non-selected country can be well approximated by the weighted average of GDP in selected countries.

Below, we focus on how we can optimally diversify chosen observed covariates. We provide detailed discussions about the potential influence of unmeasured moderators in Section 6, as it is clearer to discuss it after introducing the average-site ATE estimator in Section 5.1.

4.3.1 Illustration

Before providing the methodological details, we illustrate SPS with a simple simulation study where researchers have to choose 5 sites from 20 potential sites (see Figure 2). When researchers have only 2 variables to consider (Panel (a)), it is a relatively easy task to choose 5 sites. If one can select sites close to the center of the distribution and four corners in the panel, other



(a) (b)

Figure 2: **Illustration with Simple Simulated Data.**
Note: Red circles represent values of site-level variables $\mathbf{X}$ of selected sites, while black circles represent those of non-selected sites. Panel (a) considers two variables (X_1, X_2). Panel (b) considers four variables (X_1, X_2, X_3, X_4), and each of the six plots in the middle visualizes different pairs of variables. Histograms in the last rows represent the marginal distributions of each variable for selected (red) and non-selected (black) sites.

non-selected sites will be "inside" (formally, within a convex hull) of selected sites, which means those non-selected sites can be well approximated by the weighted average of selected sites (King and Zeng, 2006). In this simple example, SPS selected sites as such (selected sites are represented by red circles and non-selected sites by black circles). Importantly, selected sites cover a wide range of values in both contextual factors (X_1, X_2). Histograms in the last row of Panel-(a) show the marginal distributions of each variable.

However, in practice, researchers often want to consider many variables $\mathbf{X}$ that are predictive of across-site heterogeneity of causal effects. As shown in Panel (b), even when they have only 4 variables, they have to simultaneously consider 6 two-dimensional figures and select sites. This task will become even more infeasible when researchers have more variables to consider. In such scenarios, the value of SPS becomes even clearer. By solving an internal optimization problem, SPS can simultaneously consider many variables and choose diverse sites. Panel (b) shows that SPS indeed selects diverse sites such that many non-selected sites are "inside" of selected sites and can be well approximated in all dimensions. As a result, selected sites could cover a wide range of values in all four contextual factors (see the last row of Panel (b)). In Section 7, we use a multi-country survey experiment to illustrate how SPS works in practice (e.g., see Figure 3).

4.3.2 Optimization Problem behind SPS

To formally introduce SPS, we require some notations. Define S_k to be a binary variable taking 1 if site k is selected as a study site and taking 0 otherwise. Thus, $\mathbf{S} = (S_1, S_2, \ldots, S_N)$ represents which sites are selected for experiments. We use W_{jk} to denote weight we assign to selected site j when predicting the ATE of non-selected site k. Then, we can define imbalance measure $B_{k\ell}$ for non-selected site k's variable ℓ as

$$B_{k\ell}(\mathbf{W}, \mathbf{S}) \;\coloneqq\; (X_{k\ell} - \sum_{j:S_j=1} W_{jk} X_{j\ell})^2,$$

which captures how well ℓth covariate of non-selected site k is approximated by the weighted average of selected sites. For example, when Germany was not selected, this $B_{k\ell}(\mathbf{W}, \mathbf{S})$ could measure how well Germany's GDP is approximated by the weighted average GDP of selected countries. Importantly, this measure is a function of both weights $\mathbf{W}$ and site-selection $\mathbf{S}$.

SPS minimizes the average imbalance among non-selected sites by selecting optimal sites $\mathbf{S}$ and weights $\mathbf{W}$. For presentational clarity, we start with the most basic version of SPS below and then later provide a recommended version that builds on the following basic one. Formally, the basic version of SPS solves the following minimization problem.

$$\min_{(\mathbf{S},\, \mathbf{W})} \quad \frac{1}{N - N_S} \sum_{k=1}^{N} \underbrace{(1 - S_k)\left(\frac{1}{L} \sum_{\ell=1}^{L} B_{k\ell}(\mathbf{W}, \mathbf{S})\right)}_{\text{Imbalance for non-selected site } k} \tag{2}$$

with standard constraints that (i) the number of selected sites is N_S ($\sum_{k=1}^{N} S_k = N_S$), and (ii) weights are positive and sum to one ($\mathbf{W} \geq 0$, $\sum_{j:S_j=1} W_{jk} = 1$ for non-selected sites k). In practice, we standardize each variable to make the scale of variables comparable.

By solving the optimization problem, we get two outputs at the same time. First, we get the optimal selection of sites $\widehat{\mathbf{S}}$. These sites are selected such that non-selected sites can be well approximated by the selected sites. Second, we also get weights $\widehat{\mathbf{W}}$ that we use to approximate non-selected sites using the selected sites.

The objective function consists of two parts. First, $\frac{1}{L}\sum_{\ell=1}^{L} B_{k\ell}(\mathbf{W}, \mathbf{S})$ captures the imbalance for site k, averaging over L site-level covariates. Second, by multiplying this average imbalance by $(1 - S_k)$, the objective function averages over the imbalance only for non-selected sites k with $S_k = 0$. Overall, the objective function represents how well the site-level variables $\mathbf{X}$ of non-selected sites $\mathcal{R}$ are approximated by the weighted average of selected sites $\mathcal{S}$. SPS minimizes this overall imbalance, thus finding the selection of sites and weights that make this approximation the best. In Section 5.3, we provide a more formal justification for the proposed objective function by considering the mean squared error of downstream analyses.

Note that when sites were already selected (i.e., $\mathbf{S}$ is fixed), one only needs to estimate weights, which is similar to the optimization problem of the original synthetic control method (Abadie *et al.*, 2010).[4] When one wants to consider internal validity and choose treatment assignment, this optimization is similar to the synthetic design (Abadie and Zhao, 2021; Doudchenko *et al.*, 2021). The main difference is that SPS selects sites for external validity (rather than treatments for internal validity), which will lead to different causal estimands, types of constraints we add to the main optimization problem, and downstream causal estimators.

4.3.3 Incorporating Domain Knowledge and Practical Constraints into SPS

In practice, we recommend incorporating additional constraints informed by practical considerations and substantive theories of interest. Table 2 summarizes examples of domain knowledge and practical constraints users may add to SPS.

First, researchers can easily incorporate logistical and ethical constraints. For example, when a survey firm does not have an online panel in China, users can add $S_{\text{China}} = 0$ as a constraint, which guarantees that China will not be selected, and SPS will select other diverse sites and weights within this constraint. Similarly, if users want to always select Uganda as one of the study sites because it is a hard test for a given theory and they have a long-standing local partner, they can add $S_{\text{Uganda}} = 1$ as a constraint.

Second, as currently done in conventional purposive sampling, it is recommended to stratify SPS to prioritize important site-level variables. For example, users can make sure to have at least 2 democracies and at least 2 autocracies by adding $\sum_{k:S_k=1} \text{Dem}_k \geq 2$ and $\sum_{k:S_k=1} \text{Auto}_k \geq 2$ as constraints where Dem_k and Auto_k are equal to one when site k is democracy and autocracy, respectively. If users are worried about selecting too many extreme cases, they can explicitly stratify the SPS algorithm to choose both typical and diverse sites. For example, researchers can make sure to select at least one country from each tercile of GDP by adding $\sum_{k:S_k=1} \mathbf{1}\{\text{GDP}_k \leq 33 \text{ percentile}\} \geq 1$, $\sum_{k:S_k=1} \mathbf{1}\{\text{GDP}_k \in (33 \text{ and } 66 \text{ percentiles})\} \geq 1$, and $\sum_{k:S_k=1} \mathbf{1}\{\text{GDP}_k \geq 66 \text{ percentile}\} \geq 1$ as constraints where $\mathbf{1}\{\cdot\}$ is an indicator function and $\mathbf{1}\{\text{GDP}_k \in (33 \text{ and } 66 \text{ percentiles})\}$ are equal to one only when GDP of site k is within 33

[4]While our method is inspired by the synthetic control method in the sense that we use the weighted averages of selected sites to approximate non-selected sites, our method does not presume data on a history of pre-treatment outcome variables, which is one of the most important features of the original synthetic control method. As a result, instead of a common panel data framework, we use a different type of statistical framework, which we discuss further in Section 5.3.

Examples	**Formalization** (Add constraints to SPS)
Practical Constraints	
We cannot conduct studies in site k e.g., No online survey firm in China	$S_k = 0$ for infeasible site k
Domain Knowledge	
We always want to select site k e.g., Select Uganda because it is a hard test	$S_k = 1$ for site k we always select
Stratification	
We want to select studies from different groups e.g., Select at least 2 democracies and 2 autocracies	$\sum_{k:S_k=1} \text{Dem}_k \geq 2$ and $\sum_{k:S_k=1} \text{Auto}_k \geq 2$
We want to select both typical sites and diverse sites e.g., Select one country from each tercile of GDP	$\sum_{k:S_k=1} \mathbf{1}\{\text{GDP}_k \leq 33 \text{ percentile}\} \geq 1$ $\sum_{k:S_k=1} \mathbf{1}\{\text{GDP}_k \in (33 \text{ and } 66 \text{ percentiles})\} \geq 1$ $\sum_{k:S_k=1} \mathbf{1}\{\text{GDP}_k \geq 66 \text{ percentile}\} \geq 1$

Table 2: **Common Examples of Practical Constraints and Domain Knowledge that Users can Incorporate into SPS.** *Note:* The companion `R` package `spsR` can help incorporate these constraints using simple functions.

and 66 percentiles of the GDP distribution. Finally, there are many other domain knowledge researchers can add to the proposed SPS, e.g., budget constraints, differential costs of each site, and differential importance of each site-level variable.

Users can also add penalty terms to improve the basic SPS algorithm. First, to avoid relying on extreme cases, users can add the following penalty term to prioritize sites closer to non-selected sites.

$$\frac{1}{N - N_S} \sum_{j=1}^{N} \sum_{k=1}^{N} W_{jk} \underbrace{S_j(1 - S_k) \frac{1}{L} \sum_{\ell=1}^{L} (X_{j\ell} - X_{k\ell})^2}_{\substack{\text{Distance between} \\ \text{Selected Site } j \text{ and Non-Selected Site } k}}, \tag{3}$$

which captures the weighted average of the pair-wise distance between selected site j and non-selected site k. By incorporating this as the penalty term, users can make SPS more robust to outliers. As discussed above, we also recommend using simple stratification if users are worried about extreme cases. Second, users can also add the following penalty term to encourage uniform weights, which will increase efficiency of estimating the average-site ATE.

$$\frac{1}{N - N_S} \sum_{j=1}^{N} \sum_{k=1}^{N} S_j(1 - S_k) W_{jk}^2 \tag{4}$$

We provide more formal discussions about these penalty terms in Section 5.3.

5 From Site Selection to External Validity Analysis

Once we complete studies in each selected site, how can we aggregate evidence for external validity analysis? In this section, we consider how to estimate the average-site ATE by combining causal estimates from selected sites. After introducing the proposed SPS estimator (Section 5.1), we discuss connections to and differences from meta-analysis estimators popular in the social sciences (Section 5.2). Finally, we describe the theoretical property of the SPS estimator (Section 5.3).

5.1 SPS Estimator

After selecting sites and conducting experiments in those selected sites, researchers can use the conventional ATE estimator $\widehat{\theta}_j$, e.g., difference-in-means, for selected sites $j \in \mathcal{S}$. If researchers use quasi-experimental observational studies, they can also use existing estimators for $\widehat{\theta}_j$ under corresponding identification assumptions.

For non-selected sites $k \in \mathcal{R}$, researchers can use the following weighted average estimator.

$$\widehat{\theta}_k^W \coloneqq \sum_{j \in \mathcal{S}} \widehat{W}_{jk} \widehat{\theta}_j$$

where weights $\widehat{W}_{jk}$ are estimated in SPS (equation (2)). Importantly, these weights are estimated such that this non-selected site k is well approximated by the weighted average of the selected sites.

The proposed SPS estimator for the average-site ATE is then defined as,

$$\widehat{\theta}_{AS} \coloneqq \frac{1}{N} \Big(\sum_{j \in \mathcal{S}} \widehat{\theta}_j + \sum_{k \in \mathcal{R}} \widehat{\theta}_k^W \Big) \tag{5}$$

where $\mathcal{S}$ denotes all the selected sites and $\mathcal{R}$ denotes all the non-selected sites. This simply averages over the site-specific ATE estimates from selected and non-selected sites. We show that this SPS estimator can be rewritten as the weighted average of the ATEs in the selected sites.

$$\widehat{\theta}_{AS} = \sum_{j \in \mathcal{S}} \widetilde{W}_j \widehat{\theta}_j, \qquad \text{where } \widetilde{W}_j = (1 + \sum_{k \in \mathcal{R}} \widehat{W}_{jk})/N \text{ and } \sum_{j \in \mathcal{S}} \widetilde{W}_j = 1.$$

We provide proof of this equivalence and propose the conservative variance estimator in Appendix B.3.

We summarize several advantages of the proposed estimator. The first is its simplicity and interpretability. The final estimator is a simple weighted average of the ATEs in the selected sites. Therefore, researchers can see exactly how estimates for each non-selected site k and the average-site ATE were created. We inherit this desirable property from the synthetic control

method, even though the exact use case of our method (site selection for external validity) is different from the one for the original synthetic control method. Because the SPS estimator is a weighted average estimator, it also inherits desirable properties of weighted average estimators — the SPS estimator is a safe and conservative estimator by avoiding extrapolation, and it is a stable estimator that works well with small sample sizes.

The proposed SPS estimator is the optimal weighted average-based predictor that minimizes the worst-case mean squared error (we provide formal, technical discussions in Section 5.3). This theoretical foundation has several key practical implications. First, we only view the SPS estimator to be an optimal predictor given observed site-level variables, and importantly, we do not view the SPS to be an unbiased estimator given the possibility of unobserved moderators. Due to the inherent difficulty of external validity analysis, it is often impossible to obtain an unbiased estimate of the average site ATE without (often infeasible) random sampling, unless researchers make stringent modeling assumptions that we avoid in this paper. Rather, we focused on constructing an estimator that can minimize the prediction error, while explicitly allowing for unobserved moderators.

Second, because of this theoretical foundation, researchers can empirically assess the potential influence of unobserved moderators after experiments by a procedure similar to cross-validation. In particular, users can randomly choose half of the selected sites as if they were unobserved non-selected sites and predict the average ATE of those non-selected sites based on the remaining selected sites. By repeating the same procedure many times, researchers can empirically check how well the SPS estimator can credibly infer the ATEs in non-selected sites. When we pass this test, there is no evidence of significant bias from unobserved site-level variables, while we can never confirm it as in usual statistical tests. When we fail, it implies large across-site heterogeneity, not explained by site-level variables. We view this as an opportunity for further research (rather than a failure of the study) because it shows that there remains a large amount of across-site heterogeneity that existing theories cannot account for. In such scenarios, researchers can consider sequential learning: rather than viewing the current study as the final confirmation, researchers could suggest a new study by sequentially applying SPS (see our discussion in Section 8.3).

5.2 Connections to and Differences from Meta-Analysis Estimators

The proposed SPS estimator is strongly connected to typical meta-analysis estimators (e.g., Cooper *et al.*, 2019; Dunning *et al.*, 2019). The two most popular estimators in the social sciences — fixed effect and random effect meta-analysis estimators — are both weighted average estimators (e.g., Gerber and Green, 2012; Miratrix *et al.*, 2021). Thus, for those who

have used meta-analysis estimators to analyze multi-site experiments and multi-context observational studies, applying the proposed method does not introduce additional methodological or computational complications.

However, our method differs from the typical meta-analysis estimators in weight construction in a fundamental way. One of the main challenges of typical meta-analysis estimators is that they assume across-site differences in the ATEs are zero or random, and, thus, weights used in such meta-analysis estimators do not take into account systematic differences across sites. As a result, these weights are appropriate only when sites are randomly sampled from a population of sites, which is rarely the case in social science applications, as we show in our literature review. In contrast, our SPS estimator explicitly takes into account site-level differences in terms of user-specified covariates $\mathbf{X}$, and weights are estimated such that covariates of non-selected sites are well approximated by the weighted average of covariates of selected sites. Thus, our SPS estimator allows researchers to take into account systematic differences across sites, while using a familiar weighted average estimator.

Note that meta-regression is an alternative popular meta-analysis estimator to take into account systematic differences across sites. However, its reliable use often requires a much larger number of sites (e.g., in psychology, education, and medicine, where meta-regression is more popular, the average number of included studies is about 65 (Tipton *et al.*, 2019)), and thus, it has been infeasible in most multi-site studies in political science and other social science fields. When there are a large number of sites, our SPS estimator and meta-regression are complementary to each other.

5.3 Formal Properties

In this section, we show that, within a large class of weighted average estimators, the SPS estimator minimizes the worst-case mean squared error.

First, we can define the mean squared error of any weighted average estimator as follows.

$$\text{MSE} \coloneqq \frac{1}{N - N_S} \sum_{k \in \mathcal{R}} \mathbb{E}\left\{ \left(\theta_k - \widehat{\theta}_k(\mathbf{W}) \right)^2 \right\}$$

where $\widehat{\theta}_k(\mathbf{W}) \coloneqq \sum_{j \in \mathcal{S}} W_{jk} \widehat{\theta}_j$ is a general weighted average estimator of the site-specific ATE in site k. This quantity takes the average of the mean squared error in each non-selected site.

At the site selection stage, because we have not yet collected data, we cannot directly compute this mean squared error. However, we can instead examine the upper bound of the mean squared error. Formally, we show that

$$\text{MSE} \quad \lesssim \quad \lambda_1 \times \frac{1}{N - N_S} \sum_{k=1}^{N} (1 - S_k) \left(\frac{1}{L} \sum_{\ell=1}^{L} B_{k\ell}(\mathbf{W}, \mathbf{S}) \right)$$

$$+ \lambda_2 \times \frac{1}{N - N_S} \sum_{j=1}^{N} \sum_{k=1}^{N} W_{jk} S_j (1 - S_k) \frac{1}{L} \sum_{\ell=1}^{L_g} (X_{j\ell} - X_{k\ell})^2$$
$$+ \lambda_3 \times \frac{1}{N - N_S} \sum_{j=1}^{N} \sum_{k=1}^{N} S_j (1 - S_k) W_{jk}^2. \tag{6}$$

where $\lesssim$ means that the inequality holds up to some constants unrelated to $(\mathbf{S}, \mathbf{W})$. We provide proof in Appendix B.1. $(\lambda_1, \lambda_2, \lambda_3)$ are some constant parameters that capture the relative importance of the three terms. We provide the intuitive meaning of these parameters below, and we provide the analytical expression of these terms in Appendix B.1.

The first term on the right-hand side is exactly the same as the main objective function of the SPS algorithm (equation (2)) that represents how well covariates of non-selected sites can be well approximated by the weighted average of covariates in selected sites. When observed site-level variables $\mathbf{X}$ is expected to explain a larger amount of across-site heterogeneity of ATEs, this term becomes more important (λ_1 is larger). The second term is equivalent to the penalty term in equation (3) that encourages SPS to select sites closer to non-selected sites. When the underlying data generating process is expected to be highly non-linear in observed site-level variables, this term becomes more important (λ_2 is larger). Finally, the third term is equivalent to another penalty term in equation (4) that encourages uniform weights. When researchers expect that unobserved site-level variables explain a larger amount of across-site heterogeneity, this term becomes more important (λ_3 is larger). We provide technical discussions about tuning parameters $(\lambda_1, \lambda_2, \lambda_3)$ in Appendix B.1.

Because the SPS algorithm directly minimizes the right-hand side of equation (6), the SPS estimator is a minimizer of the worst-case mean squared error. This provides rigorous theoretical guarantees for the SPS algorithm in practice. It is important to emphasize two points: (1) this theoretical guarantee explicitly allows for the possibility of unobserved moderators, and (2) it avoids stringent parametric modeling assumptions.

6 Practical Guides

6.1 How to Use SPS

Here, we briefly summarize the implementation details of SPS. Our companion R package `spsR` can implement all the steps below via simple functions.

The first step is to select a population of sites of theoretical interest. This defines the target against which external validity of a given substantive theory is evaluated. Specification of the target population is essential because no causal finding is universally externally valid (Egami and Hartman, 2022); a study in a completely different context should, of course, return

a different result. This step is similar to clarifying the studies' scope conditions, and thus, this choice should be guided by a given substantive research question. As articulated in Findley *et al.* (2023), when choosing the target population of sites, it is useful to think about both Type-I external validity error (making the scope condition too broad and including sites that a given theory cannot explain) and Type-II external validity error (making the scope condition too narrow and excluding sites that a given theory can explain).

The second step is to select site-level variables that researchers want to diversify. Formally, we choose contextual moderators that predict differences in the ATEs across sites. This step is what researchers are already doing implicitly when diversifying one or two variables by hand. With SPS, researchers can incorporate all the covariates that are expected to moderate the ATEs across sites (instead of just one or two variables). At the same time, we recommend against including too many irrelevant variables because SPS might decrease the balance on key variables to improve the balance on such irrelevant variables. When there is a concern about unobserved site-level variables, it is recommended to include some proxies of such unobserved variables, if feasible. By diversifying such observed proxies that are associated with the unobserved variables, SPS can often mitigate the potential influence of unobserved variables.

After the first two steps, SPS can optimally diversify chosen site-level variables. It is recommended to stratify key site-level variables. For example, researchers can make sure to select at least one country from each region of the world. Another common scenario is that researchers want to make sure to select both typical sites and diverse sites. Users can take this into account in SPS by adding stratification such that SPS selects sites from each tercile (one from less than 33 percentile, one between 33 and 66 percentiles, and one from greater than 66 percentile), for example. This type of stratification is already common in the current practice, and it can be explicitly implemented within SPS. Researchers can also incorporate other domain knowledge and practical constraints into SPS. For example, researchers cannot run experiments in certain places, or they want to include a certain site because the site provides a hard test. Please see Section 4.3.3 for more details.

Once sites are selected and studies are completed in each selected site, researchers can first report internal validity analyses, focusing on causal estimates within selected sites. Because SPS has selected diverse sites, researchers can investigate how causal estimates vary across diverse contexts. Then, researchers can use the SPS estimator, proposed in Section 5, to estimate the average-site ATE and test whether causal findings in selected sites generalize to a broader population of sites.

Finally, to empirically test the potential influence of unobserved site-level variables, re-

searchers can conduct site-level cross-validation. In particular, users can randomly choose half of the selected sites as if they were unobserved non-selected sites and predict the average ATE of those non-selected sites based on the remaining selected sites. See Section 5.1 for more details.

6.2 Clarifications and Precautions

In this section, we provide some clarifications and precautions.

Random Sampling. If random sampling is feasible and the number of sites to be sampled is relatively large, it is recommended to rely on random sampling. However, in many social science settings, random sampling has been infeasible (recall that our literature review finds that only 2 studies could use random sampling). The proposed SPS is complementary to random sampling and is most useful when random sampling is infeasible.

How to Think about Unobserved Moderators. Researchers might be worried about the potential influence of unobserved moderators. We clarify several points about how to reason about unobserved moderators. (1) Compared to the current practice of purposive sampling, where researchers often only focus on one or two variables (recall our literature review in Appendix A.1), this concern is mitigated in SPS because users can include any number of site-level moderators based on their domain and theoretical knowledge. (2) Diversifying observed site-level variables can often help diversify even unobserved site-level variables when many key site-level variables are correlated. In addition, if unobserved variables are independent of observed site-level variables, this does not lead to unobserved bias because the distribution of unobserved variables will be the same in selected and non-selected sites, if we select sites only based on observed variables. Therefore, SPS will make the potential influence of unobserved moderators bigger only when diversifying observed site-level variables somehow *reduces* the diversity of unobserved site-level variables, which requires users to believe some complicated nonlinear relationships between observed and unobserved site-level variables. (3) While diversifying observed site-level variables can often diversify unobserved site-level variables, it is always recommended to empirically assess the influence of unobserved moderators using site-level cross-validation (see Section 5.1 and our example in Section 7). (4) Finally, SPS focuses on the mean squared error and does not assume the absence of unobserved moderators, so its theoretical guarantees are valid even if there exist unobserved moderators. The SPS algorithm can reduce the mean squared error further if users can include more predictive moderators, but unobserved moderators do not invalidate the use of SPS.

Additional Domain Knowledge. Researchers often have some domain knowledge that is not directly captured by site-level variables. For example, some sites might be of substantive

importance to a given literature, and researchers might want to prioritize such substantively important sites. Other common reasons could include sites being typical of a given theory or a hard test. One general approach to incorporate such additional information is to use stratification. For example, when choosing five sites, researchers can make sure that SPS will select at least three substantively important sites. SPS will optimally diversify site selection within such constraints.

Agnostic Use of SPS. Researchers might not be explicitly interested in estimating the average-site ATE, and they might be only interested in selecting diverse sites with transparency and flexibility. In such cases, the SPS algorithm can still be used as an agnostic site selection approach to diversify observed site-level covariates, while accommodating logistical and ethical constraints. The only difference is that, after conducting studies in each selected site, researchers will only estimate causal estimates within selected sites and compare them across diverse sites, without making explicit claims about the average-site ATE. This type of scenario might be more common when practical and ethical constraints are so severe that the target population of sites is not theoretically well motivated. Even in such cases, diversifying site selection within constraints helps users better investigate across-context heterogeneity rather than simply resorting to pure convenience sampling.

Site-Hacking. It is important to advise *against* "site-hacking," i.e., re-running SPS until researchers select sites that they wanted to select, while justifying site selection as if it were selected without any additional constraint. For example, suppose researchers only have local partners to run experiments in three unrepresentative locations, but to justify their site selection, they decide to run SPS many times until it selects the three sites and report such site selection, without clarifying their logistical constraints. SPS should not be used for such site-hacking. Importantly, this risk exists even for random sampling because researchers can re-run random sampling until they can select sites that they want. It is recommended to transparently report practical constraints (e.g., the local partner can run experiments only in three locations) and optimally diversify site selection within such constraints. In addition, researchers should always report the distribution of all relevant site-level variables in selected sites and a population of sites such that any obvious imbalance can be transparent to readers.

Meta-Regression. When conducting multi-site causal studies, another potential goal is to model heterogeneity of the ATE via site-level variables using meta-regression. While this question is crucial, it is an even more difficult problem than estimating the average-site ATE, which is already more challenging than internal validity problems. In particular, when the number of study sites is relatively small, as in political science and other social science fields (recall that

our literature review finds that the median number of study sites is 3 and the 80th percentile is 6.6), researchers have to estimate the effects of 5 site-level variables using only 6 study sites, for example. Indeed, in areas where meta-regression is more popular, e.g., psychology, education, and medicine, the number of included studies is much larger and is about 65 on average (Tipton *et al.*, 2019), whereas only 13% of multi-site experiments in political science have more than 10 sites. Given this data constraint, most applications of multi-site causal studies in political science have focused on the average-site ATE, as we do in this paper. When the number of study sites is large enough for reliable meta-regression, SPS is still a useful approach to diversify covariates, while it is not an optimal approach. For readers interested in running meta-regression, we refer readers to Tipton *et al.* (2019).

7 Empirical Application: Multi-Site Survey Experiment

In this section, we illustrate SPS using Naumann *et al.* (2018), which uses a multi-country survey experiment — one of the most common types of multi-site experiments in recent years. In particular, we will conduct empirical validation — pretending that we can only select a subset of sites that the original authors could actually study and validating whether we can recover the benchmark estimate of the average-site ATE. By doing so, we can simultaneously test the real-world performance of the method and illustrate the use of the proposed SPS step-by-step.

7.1 Background: Naumann *et al.* (2018)

In this influential study, the original authors are interested in answering a long-standing question in the immigration literature — whether and how much do natives prefer highly skilled migrants to low skilled migrants? While previous studies have conducted experiments in the US and Switzerland (e.g., Hainmueller and Hiscox, 2010; Malhotra *et al.*, 2013; Helbling and Kriesi, 2014), Naumann *et al.* (2018) tackled this question by running survey experiments in 15 European countries (Austria, Belgium, Czech Republic, Denmark, Finland, France, Germany, Ireland, Netherlands, Norway, Slovenia, Spain, Sweden, Switzerland, and the United Kingdom).

In each country, they recruited nationally representative samples and conducted a survey experiment where they randomly changed the skill level of hypothetical immigrant groups ("professionals" or "unskilled labourers") and asked respondents to show the support level for this immigrant group. They found that, in all 15 countries, the ATEs of immigrants' skills on the support level were positive and respondents preferred highly skilled immigrants to low skilled immigrants, while there are substantial variations in effect size across countries.[5] The average-

[5]The original paper reports other important substantive findings, but we focus on the ATE estimates for our empirical application. Please read the original paper for more details.

site ATE is 28.5 percentage points, and the site-specific ATE ranges from 13.5 percentage points to 41.3 percentage points.

We will use this study as a basis for our empirical validation. In particular, we will only select 6 sites out of 15 sites using SPS. Using the selected 6 sites, we will estimate the average-site ATE of all 15 sites. Because the original authors actually conducted experiments in all 15 sites, we can compare our SPS estimate based only on 6 sites to the actual experimental benchmark estimate. Given the large across-site heterogeneity, valid site selection is essential for estimation of the average-site ATE.

7.2 Site Selection

The first step of SPS is to specify the target population of sites against which we evaluate external validity. As clarified in Sections 4.2 and 6, the target population of sites should be chosen based on a given substantive theory of interest. In this application, for the sake of a clear presentation, we will use all 15 European countries as the target population of sites.

The second step is to specify site-level variables to diversify. For this, we include all 7 variables discussed in the original paper. The first three variables (GDP, size of migrant population, and unemployment rates) are country-level variables that are common in the immigration literature and are likely to explain the across-site heterogeneity of the ATEs. The "general support for immigration" variable (measured in previous waves of the European Social Survey) is a key variable that is likely to capture the baseline level of support for immigration. While SPS does not require any pre-treatment outcome variables, those variables are often quite useful for explaining across-site heterogeneity and for SPS. Finally, the last three variables (the proportion of females, the mean age, and the mean education) are country-level summary measures based on individual-level characteristics. As done here, when users want to include individual-level variables, they can include their summary measures, like means and variances, as site-level variables.

The final step is to run SPS, while taking into account stratification and any logistical constraints. In this application, as recommended in Section 6, we include stratification: for each variable, we make sure to select at least 1 site above 0.5 standard deviation and at least 1 site below -0.5 standard deviation. Users can also add any other constraints they face.

SPS selected Sweden, Norway, Spain, Switzerland, the Czech Republic, and the United Kingdom as 6 study sites. Figure 3 visualizes the results of SPS. To make visualization cleaner, we standardized each variable such that each variable has a mean zero and a standard deviation one. SPS successfully diversified each variable, covering sites with smaller values, close to the mean, and with larger values. While the last row allows users to investigate the marginal

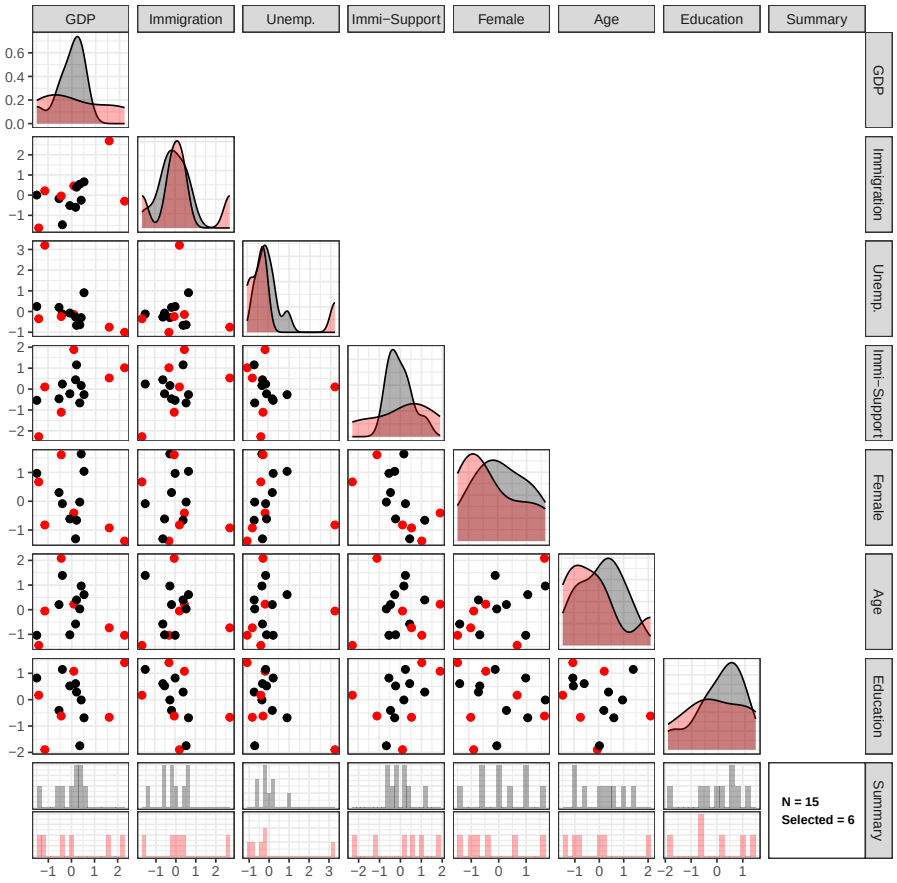
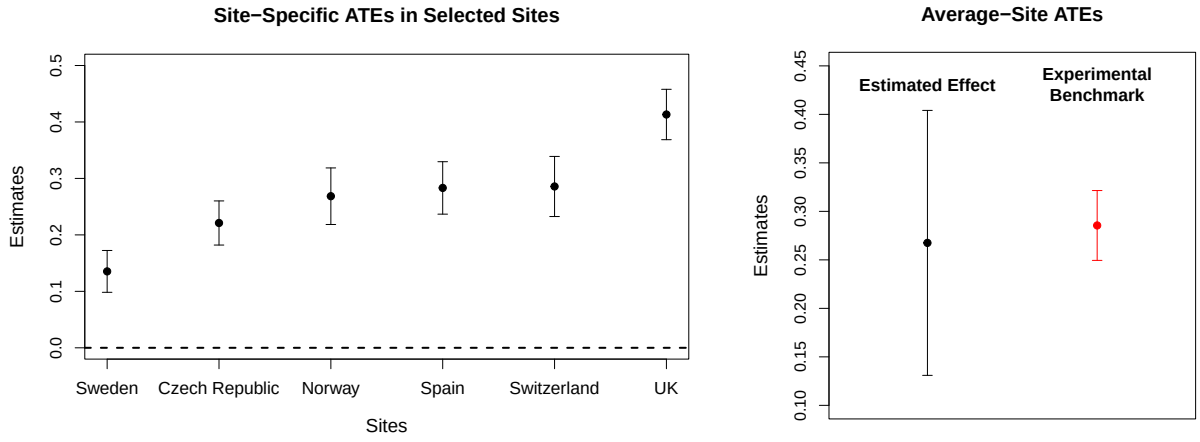


Figure 3: **Site Selection via SPS.**
Note: Red (black) points represent selected (non-selected) sites. In the last row and the diagonal plots, we can investigate the marginal distributions of each variable, and it is clear that SPS diversified all seven variables effectively. The remaining 21 figures in the middle show bivariate relationships, which also show great diversity introduced by SPS site selection.

distributions of each variable, 21 two-dimensional figures in the middle allow users to check whether selected sites could diversify bivariate relationships as well. In all 21 figures, SPS shows great diversification of site selection. While it is extremely difficult for humans to simultaneously diversify 7 variables, SPS allows users to naturally diversify all chosen variables.

We can also investigate SPS weights to understand how each non-selected country is approximated by the weighted average of the selected countries. For example, Germany is approximated with the following weights: Sweden (0.58), United Kingdom (0.16), Switzerland (0.13), Norway (0.11), Spain (0.01), and Czech Republic (0.01).



(a) (b)

Figure 4: **Panel (a): Site-Specific ATEs. Panel (b): Average-Site ATE.**
Note: The left panel reports estimated site-specific ATEs in selected sites. The right panel reports an estimate of the average-site ATE and compares it against the experimental benchmark.

7.3 External Validity Analysis

Once experiments are conducted in each site, researchers can first report site-specific ATE estimates by focusing on internal validity. Figure 4-(a) shows the results of the site-specific ATEs in the selected sites. Because SPS diversified site selection, we see large heterogeneity even across the selected sites.

For external validity analysis, we can combine causal estimates from selected sites to estimate the average-site ATE. By using the SPS estimator (equation (5)), we estimated the average-site ATE to be 26.8 percentage points (95% CI = [13.1, 40.4]). Figure 4-(b) visualizes the results. Because this is empirical validation, we can compare our estimate to the actual experimental benchmark estimated from all 15 sites. The experimental benchmark is 28.5 percentage points (95% CI = [24.9, 32.1]). Several points are worth noting. First, the point estimate from the proposed SPS is close to the experimental benchmark and is within the 95% confidence interval. Second, as expected, the standard error of the SPS estimator based only on 6 sites is much larger than that of the experimental benchmark based on 15 sites. The difference in standard errors can be interpreted as the gain from conducting experiments in more sites.

Finally, it is recommended to investigate the potential influence of unobserved moderators using site-level cross-validation. In particular, we randomly choose 3 of the selected sites as if they were unobserved non-selected sites and predict the average ATE of those 3 non-selected sites based on the remaining 3 selected sites. By repeating the same procedure many times, we can test the null hypothesis that the average ATE of those 3 non-selected sites is equal to

the SPS estimate based on the remaining 3 selected sites. We estimated the p-value to be 0.73, finding no evidence of significant bias from unobserved moderators, which is consistent with our comparison against the experimental benchmark.

8 Discussion

8.1 Connections to and Differences from Case Selection

While our main focus is on multi-site quantitative studies, this paper also has important connections to the large, influential literature on case selection in qualitative case studies (e.g., Lieberman, 2005; Gerring, 2006; Fearon and Laitin, 2008). The qualitative case selection literature has developed a wide variety of purposive sampling strategies, including typical, diverse, extreme, and deviant case selection, among others (e.g., Seawright and Gerring, 2008; Herron and Quinn, 2016). In particular, the most common practice in multi-site quantitative studies is an instance of diverse case selection, which accounts for more than 80% of the justification in multi-site experimental studies in the top political science journals. Most importantly, our proposed SPS can also be seen as a hybrid of ideas from this qualitative case selection literature (purposive diverse sampling) and from the quantitative causal inference literature (synthetic control method).

We also want to emphasize some key differences. First, SPS requires data on site-level variables that quantify key contextual factors explaining across-site heterogeneity. In qualitative case studies, researchers might utilize various knowledge that is difficult to quantify. Second, in multi-site quantitative studies that we focus on, researchers often conduct confirmatory analyses (e.g., testing hypotheses or estimating causal effects), and SPS is designed for such purposes. In contrast, in case studies, the main goal might be exploratory analyses to generate new hypotheses or theories. Finally, the goal of site selection in multi-site quantitative studies is external validity because internal validity analysis is conducted within each site. However, in some case studies, researchers compare cases to make causal, internally valid claims by using case selection methods for internal validity (e.g., most similar and most different case selection). Despite these differences, we hope that site selection methods in the quantitative literature and case selection methods in the qualitative literature can keep learning from each other.

8.2 Other Dimensions of External Validity

Even though this paper focused on the external validity question about contexts, we emphasize the importance of other dimensions of external validity, such as treatments, outcomes, and populations (Egami and Hartman, 2022). In particular, recent papers emphasize the importance of considering issues of treatments and outcomes from different angles, e.g., measurement

harmonization (see Slough and Tyson (2022a) and Wilke and Samii (2023)), the consequence of realistic and abstract treatments in survey experiments (Brutger *et al.*, 2020), the representativeness of profile distributions in conjoint analysis (de la Cuesta *et al.*, 2022), and the importance of implementation differences in field experiments (e.g., Vivalt, 2020; Angrist and Meager, 2023). As for external validity with respect to populations, many approaches allow for estimating the target population ATE under some identification assumptions (e.g., Imai *et al.*, 2008; Kern *et al.*, 2016; Dahabreh *et al.*, 2019; Egami and Hartman, 2021; Huang *et al.*, 2022), while some recent methods can assess robustness to external validity bias from non-representative populations under much weaker assumptions (e.g., Jeong and Namkoong, 2020; Gupta and Rothenhäusler, 2021; Devaux and Egami, 2022).

When useful, researchers can include information about other dimensions as site-level variables in the SPS algorithm. For example, if researchers are worried about the quality of local partners implementing the treatment, they can explicitly make sure that SPS only selects sites that have local partners with previous experiences of similar interventions.

8.3 Sequential Site Selection

While most existing multi-site studies select study sites all at once for practical reasons, it is possible to sequentially select study sites. For example, researchers can first select 3 study sites in the first phase and then select additional 3 sites in the second phase. This sequential selection can be made easily within our method because researchers can simply add a constraint that three sites are already selected in the first phase when running the SPS algorithm in the second phase. In the current practice, the number of study sites is relatively small (recall that our literature review of the top political science journals finds that the median number of study sites is 3), so the benefit of such sequential site selection might be relatively small compared to its practical cost. But we expect that, as multi-context studies will become even more important and scholars use a larger number of sites in the future, this sequential use of SPS will become more relevant.

9 Concluding Remarks

How should we select study sites for external validity? This has been a fundamental research design question for decades. As understanding the treatment assignment process is the most important question for internal validity, site selection is the most fundamental question for external validity.

For many quantitative social scientists, this question of site selection has recently become even more essential as an increasing number of scholars now use multi-site causal studies to

address external validity concerns about contexts (see Figure 1). As political science and other related social science fields pay more and more attention to external validity, we expect that this increasing trend of multi-site causal studies will continue. Another reason behind this increase is that running multi-site experiments becomes much easier both financially and practically. The cost of running randomized experiments is much lower than before, and social scientists now have much better institutional and industry support for running multi-site experiments (e.g., EGAP and JPAL provide financial, methodological, and professional training to run field experiments; many survey companies offer survey participants across the world). See Appendix A.1.1 for examples of multi-site experiments. The same trend is true for observational studies with less intensity. Observational identification strategies have become the standard in the social sciences, and scholars implement similar strategies across contexts (see Appendix A.1.2 for examples of multi-site observational studies).

Given the inherent difficulty and importance of external validity, no single approach can address all the concerns about external validity. However, we agree with many scholars that multi-site causal studies will continue to be one of the most promising, powerful strategies to address external validity concerns about contexts.

This paper offers a new methodological foundation to design such increasingly popular, reliable strategies of multi-site causal studies. We propose a general approach of synthetic purposive sampling (SPS) to select diverse sites for external validity. By combining two classical ideas from purposive sampling and the synthetic control method, we provide a simple approach to optimally select diverse sites, while taking into account logistical constraints that applied users face.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association* **105**, 490, 493–505.

Abadie, A. and Zhao, J. (2021). Synthetic Controls for Experimental Design. *arXiv preprint arXiv:2108.02196* .

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics* **130**, 3, 1117–1165.

Angrist, J. D. and Evans, W. N. (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *The American Economic Review* **88**, 3, 450–477.

Angrist, N. and Meager, R. (2023). Implementation Matters: Generalizing Treatment Effects in Education. *Available at SSRN 4487496* .

Armstrong, T. B. and Kolesár, M. (2021). Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness. *Econometrica* **89**, 3, 1141–1177.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., and Udry, C. (2015a). A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries. *Science* **348**, 6236, 1260799.

Banerjee, A., Karlan, D., and Zinman, J. (2015b). Six Randomized Evaluations of Microcredit: Introduction and Further Steps. *American Economic Journal: Applied Economics* **7**, 1, 1–21.

Bareinboim, E. and Pearl, J. (2016). Causal Inference and the Data-Fusion Problem. *Proceedings of the National Academy of Sciences* **113**, 27, 7345–7352.

Bassan-Nygate, L., Renshon, J., Weeks, J. L., and Weiss, C. M. (2023). The Generalizability of IR Experiments Beyond the US .

Ben-Michael, E., Feller, A., and Rothstein, J. (2021). The Augmented Synthetic Control Method. *Journal of the American Statistical Association* **116**, 536, 1789–1803.

Blair, G. and McClendon, G. (2020). Experiments in Multiple Contexts. In D. P. Green and J. Druckman, eds., *Handbook of Experimental Political Science*. Cambridge University Press.

Blair, G., Weinstein, J. M., Christia, F., Arias, E., *et al.* (2021). Community Policing Does Not Build Citizen Trust in Police or Reduce Crime in the Global South. *Science* **374**, 6571, eabd3446.

Brutger, R., Kertzer, J. D., Renshon, J., Tingley, D., and Weiss, C. M. (2020). Abstraction and Detail in Experimental Design. *American Journal of Political Science* .

Carnes, N. and Lupu, N. (2016). Do Voters Dislike Working-class Candidates? Voter Biases and the Descriptive Underrepresentation of the Working Class. *American Political Science Review* **110**, 4, 832–844.

Cavaille, C. and Marshall, J. (2019). Education and Anti-Immigration Attitudes: Evidence from Compulsory Schooling Reforms across Western Europe. *American Political Science Review* **113**, 1, 254–263.

Chassang, S. and Kapon, S. (2022). Designing Randomized Controlled Trials with External Validity in Mind. Tech. rep., National Bureau of Economic Research.

Choi, D. D., Poertner, M., and Sambanis, N. (2021). Linguistic Assimilation does not Reduce Discrimination against Immigrants: Evidence from Germany. *Journal of experimental political science* **8**, 3, 235–246.

Cooper, H., Hedges, L. V., and Valentine, J. C. (2019). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation.

Dahabreh, I. J., Robertson, S. E., Tchetgen Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. (2019). Generalizing Causal Inferences From Individuals In Randomized Trials to All Trial-Eligible Individuals. *Biometrics* **75**, 2, 685–694.

de la Cuesta, B., Egami, N., and Imai, K. (2022). Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution. *Political Analysis* **30**, 1, 19–45.

Dehejia, R., Pop-Eleches, C., and Samii, C. (2019). From Local to Global: External Validity in a Fertility Natural Experiment. *Journal of Business & Economic Statistics* 1–27.

DerSimonian, R. and Laird, N. (1986). Meta-Analysis in Clinical Trials. *Controlled clinical trials* **7**, 3, 177–188.

Devaux, M. and Egami, N. (2022). Quantifying Robustness to External Validity Bias. *Available at SSRN 4213753* .

Doudchenko, N., Khosravi, K., Pouget-Abadie, J., Lahaie, S., Lubin, M., Mirrokni, V., Spiess, J., and Imbens, G. (2021). Synthetic Design: An Optimization Approach to Experimental Design with Synthetic Controls. *Advances in Neural Information Processing Systems* **34**, 8691–8701.

Dunning, T., Grossman, G., Humphreys, M., Hyde, S. D., *et al.* (2019). Voter Information Campaigns and Political Accountability: Cumulative Findings from a Preregistered Meta-Analysis of Coordinated Trials. *Science Advances* **5**, 7, eaaw2612.

Egami, N. and Hartman, E. (2021). Covariate Selection for Generalizing Experimental Results: Application to a Large-Scale Development Program in Uganda. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **184**, 4, 1524–1548.

Egami, N. and Hartman, E. (2022). Elements of External Validity: Framework, Design, and Analysis. *American Political Science Review* .

Eggers, A. C., Freier, R., Grembi, V., and Nannicini, T. (2018). Regression Discontinuity Designs based on Population Thresholds: Pitfalls and Solutions. *American Journal of Political Science* **62**, 1, 210–229.

Fearon, J. D. and Laitin, D. D. (2008). Integrating Qualitative and Quantitative Methods. In H. E. Brady, J. Box-Steffensmeier, and D. Collier, eds., *The Oxford Handbook of Political Methodology*. Oxford University Press.

Findley, M. G., Denly, M., and Kikuta, K. (2023). *External Validity for Social Inquiry*.

Findley, M. G., Kikuta, K., and Denly, M. (2020). External Validity. *Annual Review of Political Science* .

Findley, M. G., Laney, B., Nielson, D. L., and Sharman, J. C. (2017). External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation. *The Journal of Politics* **79**, 3, 856–872.

Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. WW Norton.

Gerring, J. (2006). *Case Study Research: Principles and Practices*. Cambridge university press.

Green, D. P., Gerber, A. S., and Nickerson, D. W. (2003). Getting Out the Vote in Local Elections: Results from Six Door-to-Door Canvassing Experiments. *The Journal of Politics* **65**, 4, 1083–1096.

Gupta, S. and Rothenhäusler, D. (2021). The *s*-value: Evaluating Stability with respect to Distributional Shifts. *arXiv preprint arXiv:2105.03067* .

Hainmueller, J. and Hiscox, M. J. (2010). Attitudes Toward Highly Skilled and Low-Skilled Immigration: Evidence from A Survey Experiment. *American Political Science Review* **104**, 1, 61–84.

Helbling, M. and Kriesi, H. (2014). Why Citizens Prefer High-over Low-skilled Immigrants. Labor Market Competition, Welfare State, and Deservingness. *European Sociological Review* **30**, 5, 595–614.

Herron, M. C. and Quinn, K. M. (2016). A Careful Look at Modern Case Selection Methods. *Sociological Methods & Research* **45**, 3, 458–492.

Huang, M., Egami, N., Hartman, E., and Miratrix, L. (2022). Leveraging Population Outcomes to Improve the Generalization of Experimental Results. *Annals of Applied Statistics* .

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings Between Experimentalists and Observationalists About Causal Inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**, 2, 481–502.

Jeong, S. and Namkoong, H. (2020). Assessing External Validity over Worst-Case Subpopulations. *arXiv preprint arXiv:2007.02411* .

Kern, H. L., Stuart, E. A., Hill, J., and Green, D. P. (2016). Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations. *Journal of Research on Educational Effectiveness* **9**, 1, 103–127.

King, G. and Zeng, L. (2006). The Dangers of Extreme Counterfactuals. *Political Analysis* **14**, 2, 131–159.

Lieberman, E. S. (2005). Nested Analysis as A Mixed-Method Strategy for Comparative Research. *American political science review* **99**, 3, 435–452.

Lupu, Y. and Wallace, G. P. (2019). Violence, Nonviolence, and the Effects of International Human Rights Law. *American Journal of Political Science* **63**, 2, 411–426.

Magni, G. and Reynolds, A. (2021). Voter Preferences and the Political Underrepresentation of Minority Groups: Lesbian, Gay, and Transgender Candidates in Advanced Democracies. *The Journal of Politics* **83**, 4, 1199–1215.

Malhotra, N., Margalit, Y., and Mo, C. H. (2013). Economic Explanations for Opposition to Immigration: Distinguishing between Prevalence and Conditional Impact. *American Journal of Political Science* **57**, 2, 391–410.

Miratrix, L. W., Weiss, M. J., and Henderson, B. (2021). An Applied Researcher's Guide to Estimating Effects from Multisite Individually Randomized Trials: Estimands, Estimators, and Estimates. *Journal of Research on Educational Effectiveness* **14**, 1, 270–308.

Munger, K. (2019). Knowledge Decays: Temporal Validity and SocialScience in a Changing World. *Working Paper* .

Naumann, E., F. Stoetzer, L., and Pietrantuono, G. (2018). Attitudes towards Highly Skilled and Low-Skilled Immigration in Europe: A Survey Experiment in 15 European countries. *European Journal of Political Research* **57**, 4, 1009–1030.

Olsen, R. B., Orr, L. L., Bell, S. H., and Stuart, E. A. (2013). External Validity in Policy Evaluations that Choose Sites Purposively. *Journal of Policy Analysis and Management* **32**, 1, 107–121.

Raudenbush, S. W. and Liu, X. (2000). Statistical Power and Optimal Design for Multisite Randomized Trials. *Psychological methods* **5**, 2, 199.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Seawright, J. and Gerring, J. (2008). Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options. *Political research quarterly* **61**, 2, 294–308.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Slough, T., Rubenson, D., Levy, R., *et al.* (2021). Adoption of Community Monitoring Improves Common Pool Resource Management across Contexts. *Proceedings of the National Academy of Sciences* **118**, 29, e2015367118.

Slough, T. and Tyson, S. A. (2022a). External Validity and Meta-Analysis. *American Journal of Political Science* .

Slough, T. and Tyson, S. A. (2022b). Sign-Congruence, External Validity, and Replication.

Tipton, E. (2013). Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations from Experiments. *Evaluation review* **37**, 2, 109–139.

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., and Caverly, S. (2014). Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling. *Journal of Research on Educational Effectiveness* **7**, 1, 114–135.

Tipton, E. and Peck, L. R. (2017). A Design-based Approach to Improve External Validity in Welfare Policy Evaluations. *Evaluation review* **41**, 4, 326–356.

Tipton, E., Pustejovsky, J. E., and Ahmadi, H. (2019). Current Practices in Meta-Regression in Psychology, Education, and Medicine. *Research Synthesis Methods* **10**, 2, 180–194.

Tomz, M. R. and Weeks, J. L. (2013). Public Opinion and the Democratic Peace. *American Political Science Review* **107**, 4, 849–865.

Valentino, N. A., Soroka, S. N., Iyengar, S., Aalberg, T., Duch, R., *et al.* (2019). Economic and Cultural Drivers of Immigrant Support Worldwide. *British Journal of Political Science* **49**, 4, 1201–1226.

Valenzuela, A. A. and Michelson, M. R. (2016). Turnout, Status, and Identity: Mobilizing Latinos to vote with Group Appeals. *American Political Science Review* **110**, 4, 615–630.

Vivalt, E. (2020). How Much Can We Generalize from Impact Evaluations? *Journal of the European Economic Association* **18**, 6, 3045–3089.

Westreich, D., Edwards, J. K., Lesko, C. R., Cole, S. R., and Stuart, E. A. (2019). Target Validity and The Hierarchy of Study Designs. *American Journal of Epidemiology* **188**, 2, 438–443.

Wilke, A. and Humphreys, M. (2020). Field Experiments, Theory, and External Validity. In L. Curini and R. Franzese, eds., *The SAGE Handbook of Research Methods in Political Science and International Relations*. Transaction Publishers.

Wilke, A. and Samii, C. (2023). To Harmonize or Not? Research Design for Cross-Context Learning .

Online Supplementary Appendix

A Literature Review of Multi-Site Causal Studies

A.1 Literature Review Procedure

To evaluate the current practice of multi-site research, we conducted a review of academic articles published in the top 10 political science journals: American Political Science Review (APSR), American Journal of Political Science (AJPS), Journal of Politics (JOP), Political Behavior (PB), Quarterly Journal of Political Science (QJPS), British Journal of Political Science (BJPS), Comparative Political Studies (CPS), World Politics (WP), International Organization (IO), and Journal of Experimental Political Science (JEPS). These journals represent a group of highly cited and influential journals in political science. For example, these 10 journals together have total citations of over 7,800 on average as compared to the 1,315 average total citation counts across all academic journals in the field of political science. Furthermore, the 5-year journal impact factor among these 10 journals is 5.8 on average, more than twice as large as the average score across all political science journals.[6]

A.1.1 Multi-Site Experiments

To assess the current practice of multi-site experimental studies, we first searched for all articles published in the years 2000 through 2022 (inclusive) using a keyword "experiment" in Web of Science, which returned a total of 1,335 articles. We then classified whether the experiment discussed in each article is a multi-site study using the following steps. First, we used GPT to label each article as a multi-site experimental study based on the article abstract.[7] We then manually verified each of such GPT labels.

For the GPT labels, we used the following prompt to classify the experiment type, geographic unit and number of experimental sites in each article:

```
You will be provided with a summary of experimental research delimited  by
triple quotes.

Perform the following tasks:
1 - Determine the type of experiment as 'field', 'survey', 'conjoint',
'laboratory', 'natural', 'monte carlo', or 'list'. If the research does not
involve an experiment, print 'no experiment'.
2 - Print a total number of locations where the experiment was conducted.
3 - Determine the geographic unit of locations where the experiment was
conducted.
```

[6] These values are based on a total of 307 political science journals recorded in the Journal Citation Reports provided by Web of Science.

[7] We used GPT-3.5-Turbo API with zero temperature.

```
4 - List the locations where the experiment was conducted.

For all tasks, if you are not sure, print 'Unsure'.

Summarize the answers in a json object that contains the following keys:
type, num_sites, site_level, list_sites.

RESEARCH SUMMARY: ```{abstract}```

ANSWER:
```

To increase accuracy, we used multi-shot prompting by inserting six abstracts and corresponding answers prior to providing an abstract of interest. As a result, GPT classified a total of 146 articles as a multi-site experiment.

We then manually verified the 146 articles that were labeled as a multi-site by GPT as well as a random selection of 139 articles that were labeled as a non-multi-site by GPT, which returned a total of 103 articles as a multi-site out of the 285 articles reviewed: 96 out of the 146 articles labeled as a multi-site experiment by GPT were verified as such, and 7 out of the 139 articles labeled as a non-multi-site experiments by GPT were verified as a multi-site by our manual correction.

Importantly, all studies we review below are manually verified to be multi-site experiments. This means that the number of multi-site experiments we report is the lower bound.

Tables OA-1 through OA-3 show a full list of articles that conduct multi-site experiments in field, survey, and laboratory settings, respectively. Note that articles may be listed more than once if multiple types of experiments were conducted (e.g., Findley *et al.* (2017) conduct multi-site field and survey experiments).

Table OA-1: Multi-site Survey Experiments

N	Author (Year)	Journal	Title
1	Bruter (2009)	CPS	Time Bomb? The Dynamic Effect of News and Symbols on the Political Identity of European Citizens
2	Turgeon (2009)	PB	'Just Thinking:' Attitude Development, Public Opinion, and Political Representation
3	Johns and Davies (2012)	JOP	Democratic Peace or Clash of Civilizations? Target States and Support for War in Britain and the United States
4	Lu, Scheve and Slaughter (2012)	AJPS	Inequity Aversion and the International Distribution of Trade Protection
5	Lyall, Blair and Imai (2013)	APSR	Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan
6	Tomz and Weeks (2013)	APSR	Public Opinion and the Democratic Peace

Table OA-1: Multi-site Survey Experiments

N	Author (Year)	Journal	Title
7	Aaroe and Petersen (2014)	JOP	Crowding Out Culture: Scandinavians and Americans Agree on Social Welfare in the Face of Deservingness Cues
8	Ocantos, Jonge and Nickerson (2014)	AJPS	The Conditionality of Vote-Buying Norms: Experimental Evidence from Latin America
9	Jonge and Nickerson (2014)	PB	Artificial Inflation or Deflation? Assessing the Item Count Technique in Comparative Surveys
10	Bloom, Arikan and Courtemanche (2015)	APSR	Religious Social Identity, Religious Belief, and Anti-Immigration Sentiment
11	Lyall, Shiraito and Imai (2015)	JOP	Coethnic Bias and Wartime Informing
12	Carnes and Lupu (2016)	APSR	Do Voters Dislike Working-Class Candidates? Voter Biases and the Descriptive Underrepresentation of the Working Class
13	Lu and Scheve (2016)	CPS	Self-Centered Inequity Aversion and the Mass Politics of Taxation
14	Zink and Dawes (2016)	PB	The Dead Hand of the Past? Toward an Understanding of Constitutional Veneration
15	Bechtel and Scheve (2017)	JEPS	Who Cooperates? Reciprocity and the Causal Effect of Expected Cooperation in Representative Samples
16	Findley et al. (2017)	JOP	External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation
17	Gschwend, Meffert and Stoetzer (2017)	JOP	Weighting Parties and Coalitions: How Coalition Signals Influence Voting Behavior
18	Laustsen (2017)	PB	Choosing the Right Candidate: Observational and Experimental Evidence that Conservatives and Liberals Prefer Powerful and Warm Candidate Personalities, Respectively
19	Soroka et al. (2017)	JEPS	Ethnoreligious Identity, Immigration, and Redistribution
20	Wright et al. (2017)	CPS	Multiculturalism and Muslim Accommodation: Policy and Predisposition Across Three Political Contexts
21	Auerbach and Thachil (2018)	APSR	How Clients Select Brokers: Competition and Choice in India's Slums
22	Carlin and Love (2018)	BJPS	Political Competition, Partisanship and Interpersonal Trust in Electoral Democracies
23	Lee (2019)	CPS	The Revival of Charisma: Experimental Evidence From Argentina and Venezuela
24	Frye, Reuter and Szakonyi (2019)	WP	Vote Brokers, Clientelist Appeals, and Voter Turnout: Evidence from Russia and Venezuela
25	Lupu and Wallace (2019)	AJPS	Violence, Nonviolence, and the Effects of International Human Rights Law

Table OA-1: Multi-site Survey Experiments

N	Author (Year)	Journal	Title
26	Kenan and Zohlnhoefer (2019)	PB	Policy and Blame Attribution: Citizens' Preferences, Policy Reputations, and Policy Surprises
27	Valentino et al. (2019)	BJPS	Economic and Cultural Drivers of Immigrant Support Worldwide
28	Chen and MacDonald (2020)	JEPS	Bread and Circuses: Sports and Public Opinion in China
29	Chilton et al. (2020)	BJPS	Reciprocity and Public Opposition to Foreign Direct Investment
30	Goerres, Karlsen and Kumlin (2020)	BJPS	What Makes People Worry about the Welfare State? A Three-Country Experiment
31	Jensen and Rosas (2020)	JEPS	Open for Politics? Globalization, Economic Growth, and Responsibility Attribution
32	Mutz and Lee (2020)	APSR	How Much is One American Worth? How Competition Affects Trade Preferences
33	Tomz and Weeks (2020)	JOP	Human Rights and Public Support for War
34	Tomz, Weeks and Milo (2020)	IO	Public Opinion and Decisions About Military Force in Democracies
35	Avdagic and Savage (2021)	BJPS	Negativity Bias: The Impact of Framing of Immigration on Welfare State Support in Germany, Sweden and the UK
36	Pereira (2021)	PB	Do Female Politicians Face Stronger Backlash for Corruption Allegations? Evidence from Survey-Experiments in Brazil and Mexico
37	Blais and Vallve (2021)	PB	Conformity and Individuals' Response to Information About Aggregate Turnout
38	Bush and Zetterberg (2021)	AJPS	Gender Quotas and International Reputation
39	Dellmuth and Tallberg (2021)	BJPS	Elite Communication and the Popular Legitimacy of International Organizations
40	Doces and Wolaver (2021)	PB	Are WeAllPredictably Irrational? An Experimental Analysis
41	Edwards and Arnon (2021)	BJPS	Violence on Many Sides: Framing Effects on Protest and Support for Repression
42	Freire, Mignozzetti and Skarbek (2021)	JEPS	Institutional Design and Elite Support for Climate Policies: Evidence from Latin American Countries
43	Goodman (2021)	CPS	Immigration Threat, Partisanship, and Democratic Citizenship: Evidence from the US, UK, and Germany
44	Hubscher, Sattler and Wagner (2021)	BJPS	Voter Responses to Fiscal Austerity
45	Incerti et al. (2021)	BJPS	Hawkish Partisans: How Political Parties Shape Nationalist Conflicts in China and Japan
46	Kitagawa and Chu (2021)	WP	The Impact of Political Apologies on Public Opinion
47	Klasnja, Lupu and Tucker (2021)	JEPS	When Do Voters Sanction Corrupt Politicians?

Table OA-1: Multi-site Survey Experiments

N	Author (Year)	Journal	Title
48	Magni and Reynolds (2021)	JOP	Voter Preferences and the Political Underrepresentation of Minority Groups: Lesbian, Gay, and Transgender Candidates in Advanced Democracies
49	Robison et al. (2021)	CPS	Does Class-Based Campaigning Work? How Working Class Appeals Attract and Polarize Voters
50	Wood, Hoy and Pryke (2021)	JEPS	The Effect of Geostrategic Competition on Public Attitudes to Aid
51	Yu et al. (2021)	PB	The (Null) Effects of Happiness on Affective Polarization, Conspiracy Endorsement, and Deep Fake Recognition: Evidence from Five Survey Experiments in Three Countries
52	Aarslew (2022)	BJPS	Why Don't Partisans Sanction Electoral Malpractice?
53	Arias and Blair (2022)	JOP	Changing Tides: Public Attitudes on Climate Migration
54	Bayram and Graham (2022)	JOP	Knowing How to Give: International Organization Funding Knowledge and Public Support for Aid Delivery Channels
55	McGrath et al. (2022)	CPS	Parliament, People or Technocrats? Explaining Mass Public Preferences on Delegation of Policymaking Authority
56	Bergquist et al. (2022)	BJPS	The Politics of Intersecting Crises: The Effect of the COVID-19 Pandemic on Climate Policy Preferences
57	Brutger and Guisinger (2022)	JEPS	Labor Market Volatility, Gender, and Trade Preferences
58	Carnegie and Gaikwad (2022)	WP	Public Opinion on Geopolitics and Trade Theory and Evidence
59	Duch and Gimeno (2022)	CPS	Collective Decision-Making and the Economic Vote
60	Frederiksen (2022)	APSR	Does Competence Make Citizens Tolerate Undemocratic Behavior?
61	Jurado, Leon and Walter (2022)	IO	Brexit Dilemmas: Shaping Postwithdrawal Relations with a Leaving State
62	Krishnarajan and Jensen (2022)	BJPS	When Is A Pledge A Pledge?
63	Madsen et al. (2022)	APSR	Sovereignty, Substance, and Public Support for European Courts' Human Rights Rulings
64	Magni (2022)	AJPS	Boundaries of Solidarity: Immigrants, Economic Contributions, and Welfare Attitudes
65	Magni and Reynolds (2022)	PB	The Persistence of Prejudice: Voters Strongly Penalize Candidates with HIV
66	Manekin and Mitts (2022)	APSR	Effective for Whom? Ethnic Identity and Nonviolent Resistance
67	Rehmert (2022)	PB	Party Elites' Preferences in Candidates: Evidence from a Conjoint Experiment
68	Saha and Weeks (2022)	PB	Ambitious Women: Gender and Voter Perceptions of Candidate Ambition

Table OA-1: Multi-site Survey Experiments

N	Author (Year)	Journal	Title
69	Shandler et al. (2022)	BJPS	Cyber Terrorism and Public Support for Retaliation - A Multi-Country Survey Experiment
70	Simonsen and Bonikowski (2022)	CPS	Moralizing Immigration: Political Framing, Moral Conviction, and Polarization in the United States and Denmark
71	Weinberg (2022)	CPS	Feelings of Trust, Distrust and Risky Decision-Making in Political Office. An Experimental Study With National Politicians in Three Democracies
72	Williams, Gravelle and Klar (2022)	APSR	The Competing Influence of Policy Content and Political Cues: Cross-Border Evidence from the United States and Canada
73	Williamson et al. (2022)	BJPS	Preaching Politics: How Politicization Undermines Religious Authority in the Middle East
74	Xu, Kostka and Cao (2022)	JOP	Information Control and Public Support for Social Credit Systems in China

Table OA-2: Multi-site Field Experiments

N	Author (Year)	Journal	Title
1	Green, Gerber and Nickerson (2003)	JOP	Getting out the vote in local elections: Results from six door-to-door canvassing experiments
2	Nickerson (2007)	AJPS	Quality is job one: Professional and volunteer voter mobilization calls
3	Gerber and Rogers (2009)	JOP	Descriptive Social Norms and Motivation to Vote: Everybody's Voting and so Should You
4	Michelson, Bedolla and McConnell (2009)	JOP	Heeding the Call: The Effect of Targeted Two-Round Phone Banks on Voter Turnout
5	Panagopoulos (2010)	PB	Affect, Social Pressure and Prosocial Motivation: Field Experimental Evidence of the Mobilizing Effects of Pride, Shame and Publicizing Voting Behavior
6	Broockman (2013)	AJPS	Black Politicians Are More Intrinsically Motivated to Advance Blacks' Interests: A Field Experiment Manipulating Political Incentives
7	Findley, Nielson and Sharman (2013)	IO	Using Field Experiments in International Relations: A Randomized Study of Anonymous Incorporation
8	Panagopoulos (2013)	JOP	Extrinsic Rewards, Intrinsic Motivation and Voting
9	Gift and Gift (2015)	PB	Does Politics Influence Hiring? Evidence from a Randomized Experiment

Table OA-2: Multi-site Field Experiments

N	Author (Year)	Journal	Title
10	Nickerson (2015)	JOP	Do Voter Registration Drives Increase Participation? For Whom and When?
11	Nyhan and Reifler (2015)	AJPS	The Effect of Fact-Checking on Elites: A Field Experiment on US State Legislators
12	White, Nathan and Faller (2015)	APSR	What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials
13	Valenzuela and Michelson (2016)	APSR	Turnout, Status, and Identity: Mobilizing Latinos to Vote with Group Appeals
14	Broockman and Butler (2017)	AJPS	The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication
15	Findley et al. (2017)	JOP	External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation
16	Rooij and Green (2017)	PB	Radio Public Service Announcements and Voter Participation Among Native Americans: Evidence from Two Field Experiments
17	Grossman and Michelitch (2018)	APSR	Information Dissemination, Competitive Pressure, and Politician Performance between Elections: A Field Experiment in Uganda
18	Kalla and Broockman (2020)	APSR	Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments
19	Linardi and Rudra (2020)	CPS	Globalization and Willingness to Support the Poor in Developing Countries: An Experiment in India
20	Persson et al. (2020)	JEPS	Does Deliberative Education Increase Civic Competence? Results from a Field Experiment
21	Choi, Poertner and Sambanis (2021)	JEPS	Linguistic Assimilation Does Not Reduce Discrimination Against Immigrants: Evidence from Germany
22	Harris, Kamindo and Windt (2021)	JOP	Electoral Administration in Fledgling Democracies: Experimental Evidence from Kenya
23	Magni and Leon (2021)	JEPS	Women Want an Answer! Field Experiments on Elected Officials and Gender Bias
24	Moy (2021)	JEPS	Can Social Pressure Foster Responsiveness? An Open Records Field Experiment with Mayoral Offices
25	Bennion and Nickerson (2022)	PB	Decreasing Hurdles and Increasing Registration Rates for College Students: An Online Voter Registration Systems Field Experiment
26	Goerger, Mummolo and Westwood (2022)	JEPS	Which Police Departments Want Reform? Barriers to Evidence-Based Policymaking

Table OA-3: Multi-site Lab Experiments

N	Author (Year)	Journal	Title
1	Wilking (2011)	PB	The Portability of Electoral Procedural Fairness: Evidence from Experimental Studies in China and the United States
2	Enos and Gidron (2016)	JOP	Intergroup Behavioral Strategies as Contextually Determined: Experimental Evidence from Israel
3	Vincent, Blais and Pilet (2016)	JEPS	The Electoral Sweet Spot in the Lab
4	Blais and Vallve (2021)	PB	Conformity and Individuals' Response to Information About Aggregate Turnout
5	de la Cuesta et al. (2022)	JOP	Owning It: Accountability and Citizens' Ownership over Oil, Aid, and Taxes

A.1.2 Observational Studies

For observational studies, we reviewed papers that use instrumental variables, regression discontinuity design, difference-in-differences design, or natural experiment. We first searched for all articles published in the years 2000 through 2022 (inclusive) using the following keywords: "natural experiment", "regression discontinuity", "instrument", "difference-in-difference", and "two-way fixed effects" in Web of Science, which returned a total of 375 articles. Similar to the steps taken for the experimental studies, we used GPT to classify whether an article conducts a multi-site study.[8] As a result, GPT classified a total of 62 articles as a multi-site observational study.

We then manually verified the 62 articles that GPT labeled as a multi-site as well as a random selection of 50 from the remaining articles. This returned a total of 27 articles as a multi-site out of the 112 articles reviewed. Importantly, all studies we review below are manually verified to be multi-site observational studies.

Tables OA-4, OA-5, OA-6, and OA-7 display the list of multi-site observational studies separated by the identification strategies: Difference-in-Difference, Regression Discontinuity, Instrumental Variables, and Natural Experiment, respectively.

[8]We used the same prompt as the one used in the experimental studies, except for the instruction under task 1. Task 1 for observational studies writes: `Determine the type of research analysis as 'research design', or 'empirical'. That is, if the research is about a statistical method or a particular research design, print 'research design'. If the research empirically test substantive theories, print 'empirical'.`

Table OA-4: Multi-site Observational Studies using Difference-in-Difference

N	Author (Year)	Journal	Title
1	Grossman, Pierskalla and Dean (2017)	JOP	Government Fragmentation and Public Goods Provision
2	Fonseca (2017)	AJPS	Identifying the Source of Incumbency Advantage through a Constitutional Reform
3	Singh (2019)	AJPS	Compulsory Voting and Parties' Vote-Seeking Strategies
4	Payson (2020)	APSR	The Partisan Logic of City Mobilization: Evidence from State Lobbying Disclosures
5	Safarpour et al. (2022)	PB	When Women Run, Voters Will Follow (Sometimes): Examining the Mobilizing Effect of Female Candidates in the 2014 and 2018 Midterm Elections

Table OA-5: Multi-site Observational Studies using Regression Discontinuity

N	Author (Year)	Journal	Title
1	Middleton and Green (2008)	QJPS	Do community-based voter mobilization campaigns work even in battleground states? Evaluating the effectiveness of MoveOn's 2004 outreach campaign
2	Folke, Persson and Rickne (2016)	APSR	The Primary Effect: Preference Votes and Political Promotions
3	Eggers et al. (2018)	AJPS	Regression Discontinuity Designs Based on Population Thresholds: Pitfalls and Solutions
4	Velez and Newman (2019)	AJPS	Tuning In, Not Turning Out: Evaluating the Impact of Ethnic Television on Political Participation (Publication with Expression of Concern. See vol. 63, pg. 807, 2019)
5	Holbein and Rangel (2020)	JOP	Does Voting Have Upstream and Downstream Consequences? Regression Discontinuity Tests of the Transformative Voting Hypothesis
6	Kessner and Warshaw (2020)	JOP	Politics in Forgotten Governments: The Partisan Composition of County Legislatures and County Fiscal Policies
7	Kirkland (2021)	JOP	Business Owners and Executives as Politicians: The Effect on Public Policy
8	Solodoch (2021)	IO	Regaining Control? The Political Impact of Policy Responses to Refugee Crises
9	Gordon and Yntiso (2022)	JOP	Incentive Effects of Recall Elections: Evidence from Criminal Sentencing in California Courts

Table OA-5: Multi-site Observational Studies using Regression Discontinuity

N	Author (Year)	Journal	Title
10	Olson and Stone (2022)	PB	The Incumbency Advantage in Judicial Elections: Evidence from Partisan Trial Court Elections in Six US States
11	Rau (2022)	CPS	Partisanship as Cause, Not Consequence, of Participation
12	Song (2022)	QJPS	The Rank Effect in Multimember District Elections

Table OA-6: Multi-site Observational Studies using Instrumental Variables

N	Author (Year)	Journal	Title
1	Woodberry (2012)	APSR	The Missionary Roots of Liberal Democracy
2	Schleiter and Tavits (2016)	JOP	The Electoral Benefits of Opportunistic Election Timing
3	Wimmer (2016)	CPS	Is Diversity Detrimental? Ethnic Fractionalization, Public Goods Provision, and the Historical Legacies of Stateness
4	Cavaille and Marshall (2019)	APSR	Education and Anti-Immigration Attitudes: Evidence from Compulsory Schooling Reforms across Western Europe

Table OA-7: Multi-site Observational Studies using Natural Experiment

N	Author (Year)	Journal	Title
1	Malesky and Samphantharak (2008)	QJPS	Predictable Corruption and Firm Investment: Evidence from a Natural Experiment and Survey of Cambodian Entrepreneurs
2	Dunning and Nilekani (2013)	APSR	Ethnic Quotas and Political Mobilization: Caste, Parties, and Distribution in Indian Village Councils
3	Lazarev et al. (2014)	WP	TRIAL BY FIRE A Natural Disaster's Impact on Support for the Authorities in Rural Russia
4	Bateson and Weintraub (2022)	JOP	The 2016 Election and America's Standing Abroad: Quasi-Experimental Evidence of a Trump Effect
5	Holman, Merolla and Zechmeister (2022)	APSR	The Curious Case of Theresa May and the Public That Did Not Rally: Gendered Reactions to Terrorist Attacks Can Cause Slumps Not Bumps
6	Iversen and Rehm (2022)	CPS	Information and Financialization: Credit Markets as a New Source of Inequality

A.2 Descriptive Analyses of Multi-Site Experiments

To further assess the current site selection approach in multi-site experimental studies, we hired two independent researchers to review the 103 verified multi-site experiment articles and code the following information:

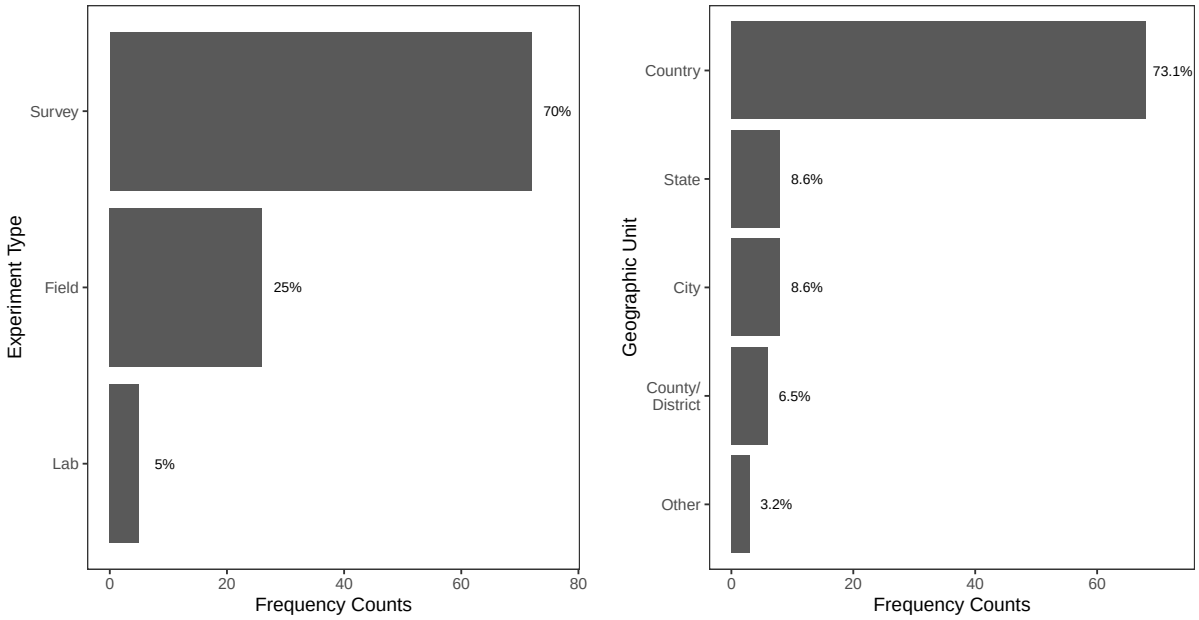
1. The geographic unit as well as the total number of experimental sites
2. Use of random sampling in selecting experimental sites
3. Use of purposive sampling in selecting experimental sites (i.e., whether the authors justify site selection by diversification of study sites)
4. Site-level variables considered when diversifying the site selection

We first show a simple count of multi-site experimental articles over time. As shown in Figure 1 in the main paper, we observe a gradual increase in a number of published articles that conduct experiments in a multi-site setting. Prior to 2010, there were only seven multi-site experiments over the span of 10 years. Since then, the number of multi-site experiments increased gradually, and most notably, the number of multi-site experiments soared to more than 20 published articles each in 2021 and 2022.

As shown in Figure OA-1 (a), we find that the majority of the multi-site experiments (70%) are survey experiments, which reflects the increasing popularity of using survey platforms such as Lucid or YouGov that possess respondents across the world. This is also reflected in panel (b) of Figure OA-1 which shows that large majority of the studies (73%) are experiments implemented in multiple countries.

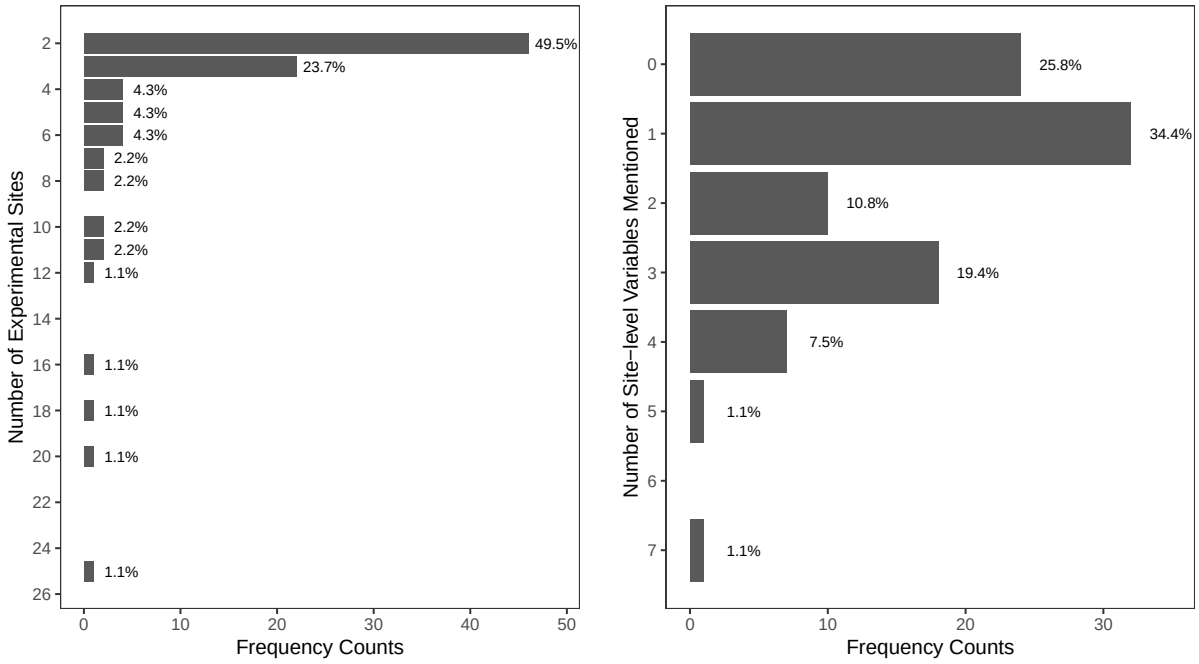
We next assess how many sites the studies conduct experiments and how the authors justify site selection. Figure OA-2 (a) shows the distribution of the number of sites from which the experiments are conducted in a given article. Over 70% of the studies we examine conduct experiments in either two or three sites, and research that conduct experiments in more than 10 locations are extremely rare.

Our record shows that almost all multi-site experimental papers reviewed select sites with purposive sampling. Only 2 articles implement a random sampling approach. When they use purposive sampling, they only mention one or two site-level variable (45.2%) to explain how the study sites vary and hence adds values to external validity.



(a) By Experimental Design (b) By Geographic Unit of Site

Figure OA-1: Breakdown of Multi-Site Experimental Studies



(a) Number of Experimental Sites (b) Number of Site-Level Variables Mentioned

Figure OA-2: Breakdown of multi-site experimental studies

B Formal Results

B.1 SPS Estimator Minimizes the Worst-Case Mean Squared Error

In this section, we clarify how SPS minimizes the worst-case mean squared error (MSE), within a large class of weighted average estimators. We consider the following general SPS algorithm.

$$\min_{(\mathbf{S}\in\{0,1\},\ \mathbf{W})} \quad \lambda_1 \times \frac{1}{N-N_S}\sum_{k=1}^{N}(1-S_k)\left(\frac{1}{L_g}\sum_{\ell=1}^{L_g}(g_\ell(\mathbf{X}_k)-\sum_{j=1}^{N}S_jW_{jk}g_\ell(\mathbf{X}_j))^2\right) \tag{OA.1}$$

$$+\ \lambda_2 \times \frac{1}{(N-N_S)}\sum_{j=1}^{N}\sum_{k=1}^{N}W_{jk}S_j(1-S_k)\frac{1}{L_g}\sum_{\ell=1}^{L_g}(g_\ell(\mathbf{X}_j)-g_\ell(\mathbf{X}_k))^2 \tag{OA.2}$$

$$+\ \lambda_3 \times \frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N}S_j(1-S_k)W_{jk}^2 \tag{OA.3}$$

$$\text{such that} \qquad \sum_{k=1}^{N}S_k=N_S,\quad \mathbf{W}\geq 0, \text{and } \sum_j S_jW_{jk}=1 \ \text{ for all non-selected sites } k \text{ with } S_k=0.$$

where $(\lambda_1,\lambda_2,\lambda_3)$ are tuning parameters. $g(\cdot)$ represents flexible transformation of site-level variables $\mathbf{X}\in\mathbb{R}^L$, such as higher-order interactions between site-level covariates and higher-order polynomials. More formally, researchers can make the transformation flexible by including basis expansion and/or kernels (relying on the theory of reproducing kernel Hilbert spaces). We use L_g to denote the dimension of $g(\mathbf{X})$ (after transformation).

In Section 4.3.2, we introduced the most basic version ($\lambda_1=1$ and $\lambda_2=\lambda_3=0$; and no transformation). As we explained there, the first part of the optimization problem (equation (OA.1)) is the most fundamental part, which makes sure that non-selected sites can be well approximated by the weighted average of the selected sites.

Two other parts (equations (OA.2) and (OA.3)) are helpful to improve the basic version of SPS, while it does not change the algorithm substantively. The second part (equation (OA.2)) acts as the penalty term for encouraging to select sites closer to non-selected sites to avoid excessive reliance on linearity on $g(\mathbf{X})$. The third part (equation (OA.3)) also acts as the penalty term for encouraging uniform weights, which will increase efficiency of the downstream weighted average estimator. These penalty terms are similar to common penalty terms in the synthetic control literature (e.g., Abadie and Zhao, 2021; Ben-Michael *et al.*, 2021; Doudchenko *et al.*, 2021).

While we provide analytical expression for $(\lambda_1,\lambda_2,\lambda_3)$ below, we summarize guiding principles here. When a linear model of $g(\mathbf{X})$ can explain a larger amount of across-site heterogeneity, λ_1 should be larger because the balance of $g(\mathbf{X})$ is crucial. When the underlying model deviates more from a linear model, λ_2 should be larger because we should select sites closer to non-selected sites to avoid excessive reliance on linearity. Finally, when unmeasured moderators have larger effects or when the variance of the site-specific ATEs in selected sites are expected to be larger, λ_3 should be larger because it is important to encourage uniform weights to reduce variance of the downstream weighted average estimator and also because site selection and

weights estimation should depend less on observed site-level variables. Please see the end of Appendix for more formal expressions of $(\lambda_1, \lambda_2, \lambda_3)$.

We consider the mean squared error (MSE) over non-selected sites, which is defined as follows.

$$\text{MSE} \coloneqq \frac{1}{N - N_S} \sum_{k \in \mathcal{R}} \mathbb{E}\left\{ \left(\theta_k - \widehat{\theta}_k^W \right)^2 \right\}$$

We show that the MSE is upper bounded by the following quantity with constant terms $(\lambda_1, \lambda_2, \lambda_3, C)$ that we define below.

$$\begin{aligned}
& \frac{1}{N - N_S} \sum_{k \in \mathcal{R}} \mathbb{E}\left\{ \left(\theta_k - \widehat{\theta}_k^W \right)^2 \right\} \\
\leq \quad & \lambda_1 \times \frac{1}{N - N_S} \sum_{k=1}^{N} (1 - S_k) \left(\frac{1}{L_g} \sum_{\ell=1}^{L_g} (g_\ell(\mathbf{X}_k) - \sum_{j=1}^{N} S_j W_{jk} g_\ell(\mathbf{X}_j))^2 \right) \\
& + \lambda_2 \times \frac{1}{(N - N_S)} \sum_{j=1}^{N} \sum_{k=1}^{N} W_{jk} S_j (1 - S_k) \frac{1}{L_g} \sum_{\ell=1}^{L_g} (g_\ell(\mathbf{X}_j) - g_\ell(\mathbf{X}_k))^2 \\
& + \lambda_3 \times \frac{1}{N - N_S} \sum_{j=1}^{N} \sum_{k=1}^{N} S_j (1 - S_k) W_{jk}^2 \quad + \quad C
\end{aligned} \tag{OA.4}$$

Our SPS algorithm (equations (OA.1)-(OA.3)) directly minimizes this worst-case MSE over site selection and weights estimation $(\mathbf{S}, \mathbf{W})$.

Proof. We introduce some notations to simplify presentation.

For selected sites $j \in \mathcal{R}$, we define $d_j \coloneqq \widehat{\theta}_j - \theta_j$. When researchers use an unbiased estimator of site-specific ATEs within each site (most common in practice), $\mathbb{E}(d_j) = 0$. As causal studies in each site use independent sets of data, d is independent across sites.

We will use the following decomposition. For all $k \in \{1, \ldots, N\}$,

$$\eta_k \coloneqq \theta_k - \{ g(\mathbf{X}_k)^\top \beta + f(g(\mathbf{X}_k)) \}$$

where η_k is a bias term of the partially linear working predictive model $g(\mathbf{X}_k)^\top \beta + f(g(\mathbf{X}_k))$ for site-specific ATE θ_k. This is a mechanical decomposition of θ_k into the bias term η_k and the working predictive model $(g(\mathbf{X}_k)^\top \beta + f(g(\mathbf{X}_k)))$, so this decomposition holds without loss of generality. We now explain each term in order. First, $g(\mathbf{X}_k)^\top \beta$ is a linear part of the working predictive model using the transformation of site-level variables $(g(\mathbf{X}_k))$ with unknown coefficients β (note that this coefficient is unknown to researchers at the site-selection stage). We assume $f(\cdot)$ is a Lipschitz function with Lipschitz constant $\rho \geq 0$, i.e., $|f(Z) - f(Z')| \leq \rho ||Z - Z'||_2$. This Lipschitz function is a large class of models (every function that is defined on an interval and has bounded first derivative is Lipschitz continuous) used widely in the literature (e.g., Armstrong and Kolesár, 2021; Ben-Michael *et al.*, 2021) and captures the deviation from linearity. Even though we allow for very flexible transformation $g(\cdot)$, it might not capture all non-linearity in observed site-level variables $\mathbf{X}_k$, and this Lipschitz function $f(\cdot)$ captures this residual non-linearity in $\mathbf{X}_k$. Finally, the working predictive model $(g(\mathbf{X}_k)^\top \beta + f(g(\mathbf{X}_k)))$ is an

extremely flexible non-linear model of observed site-level variables, but it cannot capture the influence of unobserved site-level variables, which is captured by the bias term η_k. To allow for arbitrary bias, we do not make any assumption about η_k. Importantly, θ_k is a fixed constant parameter of interest, and thus, η_k is also not random here.

To understand the MSE of weighted average estimators, we start by decomposing site-specific bias. Importantly, the following decomposition holds for any weighted average estimators.

$$
\begin{aligned}
& \theta_k - \widehat{\theta}_k^W \\
= \;& \theta_k - \sum_{j \in \mathcal{S}} W_{jk} \widehat{\theta}_j \\
= \;& \theta_k - \sum_{j \in \mathcal{S}} W_{jk} (\theta_j + d_j) \\
= \;& \left(\theta_k - \sum_{j \in \mathcal{S}} W_{jk} \theta_j \right) + \left(g(\mathbf{X}_k)^\top \beta - g(\mathbf{X}_k)^\top \beta \right) + \left(\sum_{j \in \mathcal{S}} W_{jk} g(\mathbf{X}_j)^\top \beta - \sum_{j \in \mathcal{S}} W_{jk} g(\mathbf{X}_j)^\top \beta \right) \\
& + (f(g(\mathbf{X}_k)) - f(g(\mathbf{X}_k))) + \left(\sum_{j \in \mathcal{S}} W_{jk} f(g(\mathbf{X}_j)) - \sum_{j \in \mathcal{S}} W_{jk} f(g(\mathbf{X}_j)) \right) - \sum_{j \in \mathcal{S}} W_{jk} d_j \\
= \;& \left(g(\mathbf{X}_k) - \sum_{j \in \mathcal{S}} W_{jk} g(\mathbf{X}_j) \right)^\top \beta + \left(f(g(\mathbf{X}_k)) - \sum_{j \in \mathcal{S}} W_{jk} f(g(\mathbf{X}_j)) \right) \\
& + \left(\{\theta_k - g(\mathbf{X}_k)^\top \beta - f(g(\mathbf{X}_k)\} - \sum_{j \in \mathcal{S}} W_{jk} \{\theta_j - g(\mathbf{X}_j)^\top \beta - f(g(\mathbf{X}_j))\} \right) - \sum_{j \in \mathcal{S}} W_{jk} d_j \\
= \;& \left(g(\mathbf{X}_k) - \sum_{j \in \mathcal{S}} W_{jk} g(\mathbf{X}_j) \right)^\top \beta + \left(f(g(\mathbf{X}_k)) - \sum_{j \in \mathcal{S}} W_{jk} f(g(\mathbf{X}_j)) \right) + \left(\eta_k - \sum_{j \in \mathcal{S}} W_{jk} \eta_j \right) - \sum_{j \in \mathcal{S}} W_{jk} d_j
\end{aligned}
$$

where the first line follows from the definition of a weighted average estimator, the second from the definition of d described above, the third line from add and subtract techniques (from the second to the fifth terms cancel out), the fourth from rearrangement of terms, and the final line from the definition of η_k.

To simplify notations, we now define $G_k(\mathbf{W}) \coloneqq g(\mathbf{X}_k) - \sum_{j \in \mathcal{S}} W_{jk} g(\mathbf{X}_j)$ and $F_k(\mathbf{W}) \coloneqq f(g(\mathbf{X}_k)) - \sum_{j \in \mathcal{S}} W_{jk} f(g(\mathbf{X}_j))$. Then, we have

$$
\begin{aligned}
& \mathbb{E}\left\{ \left(\theta_k - \widehat{\theta}_k^W \right)^2 \right\} \\
= \;& ||G_k(\mathbf{W})^\top \beta||_2^2 \; + \; ||F_k(\mathbf{W})||_2^2 \; + \; ||\eta_k - \sum_{j \in \mathcal{S}} W_{jk} \eta_j||_2^2 \\
& + \; 2 G_k(\mathbf{W})^\top \beta F_k(\mathbf{W}) \; + \; 2 G_k(\mathbf{W})^\top \beta \left(\eta_k - \sum_{j \in \mathcal{S}} W_{jk} \eta_j \right) \; + \; 2 F_k(\mathbf{W}) \left(\eta_k - \sum_{j \in \mathcal{S}} W_{jk} \eta_j \right) \\
& + \mathbb{E}\left\{ (\sum_{j \in \mathcal{S}} W_{jk} d_j)^2 \right\}
\end{aligned}
$$

where we used $\mathbb{E}(d_j) = 0$ and independence of d across sites.

Now we consider each term in order. We note that each bound below is not always the sharp bound (i.e., the tightest bound). As in the literature of the synthetic control method and balancing weights, we use bounds such that the resulting optimization problem has intuitive interpretation and is also computationally feasible.

For the first term, using Cauchy–Schwarz inequality,

$$||G_k(\mathbf{W})^\top \beta||_2^2 \leq ||\beta||_2^2 ||G_k(\mathbf{W})||_2^2.$$

For the second term, we obtain

$$\begin{aligned}
||F_k(\mathbf{W})||_2^2 &\coloneqq \left(f(g(\mathbf{X}_k)) - \sum_{j \in \mathcal{S}} W_{jk} f(g(\mathbf{X}_j)) \right)^2 \\
&= \left(\sum_{j \in \mathcal{S}} W_{jk} \left\{ f(g(\mathbf{X}_k)) - f(g(\mathbf{X}_j)) \right\} \right)^2 \\
&\leq \left(\rho \times \sum_{j \in \mathcal{S}} W_{jk} ||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2 \right)^2 \\
&= \rho^2 \times \sum_{j \in \mathcal{S}} W_{jk} ||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2 \sum_{j' \in \mathcal{S}} W_{j'k} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2 \\
&\leq \rho^2 \times \sum_{j \in \mathcal{S}} W_{jk} ||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2 \max_{j' \in [N]} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2 \\
&= \rho^2 \times \sum_{j \in \mathcal{S}} W_{jk} ||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2^2 \frac{\max_{j' \in [N]} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2}{||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2} \\
&\leq \rho^2 \times \frac{\max_{j' \in [N]} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2}{\min_{j' \in [N]} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2} \times \sum_{j \in \mathcal{S}} W_{jk} ||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2^2
\end{aligned}$$

where we use $[N] \coloneqq \{1, \ldots, N\}$. The first line follows from the definition, the second from rearrangement of terms, and the third from the property of the Lipschitz function $f(\cdot)$ with Lipschitz constant $\rho \geq 0$. The fourth follows from expansion of the squared term, the fifth line from $\sum_{j' \in \mathcal{S}} W_{j'k}$ being the weighted average, the sixth line from adding $||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2$, and the final line from using the minimum of the denominator.

For the third term, we obtain

$$\begin{aligned}
||\eta_k - \sum_{j \in \mathcal{S}} W_{jk} \eta_j||_2^2 &= \eta_k^2 - 2\eta_k \sum_{j \in \mathcal{S}} W_{jk} \eta_j + \sum_{j \in \mathcal{S}} W_{jk}^2 \eta_j^2 + \sum_{j \in \mathcal{S}} W_{jk} \eta_j \sum_{j' \in \mathcal{S}, j' \neq j} W_{j'k} \eta_{j'} \\
&\leq \overline{\eta}^2 + 2\overline{\eta}^2 + \overline{\eta}^2 \sum_{j \in \mathcal{S}} W_{jk}^2 + \overline{\eta}^2 \\
&= \overline{\eta}^2 \sum_{j \in \mathcal{S}} W_{jk}^2 + 4\overline{\eta}^2
\end{aligned}$$

where we use $\overline{\eta}$ to denote the unknown upper bound of $|\eta_k|$ for $k \in [N]$.

For the fifth term, we obtain

$$\begin{aligned} 2G_k(\mathbf{W})^\top \beta F_k(\mathbf{W}) &\leq 2 \times ||G_k(\mathbf{W})^\top \beta||_2 \times ||F_k(\mathbf{W})||_2 \\ &\leq 2 \times ||G_k(\mathbf{W})||_2 \times ||\beta||_2 \times \rho \times \max_{j' \in [N]} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2 \\ &\leq 2 \times ||G_k(\mathbf{W})||_2^2 \times ||\beta||_2 \times \rho \times \max_{j' \in [N]} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2 \end{aligned}$$

where the first and second lines from Hölder's inequality, Cauchy–Schwarz inequality and the bound for $||F_k(\mathbf{W})||_2$ derived above. The final line adds multiplication by $||G_k(\mathbf{W})||_2$, which can be made greater than 1 (as long as $||G_k(\mathbf{W})||_2 > 0$) by appropriately defining the scale of $g(\mathbf{X})$ and β. This final step is added for simpler interpretation because this bound can be combined together with the bound for the first term.

For the sixth term, we obtain

$$\begin{aligned} 2G_k(\mathbf{W})^\top \beta \left(\eta_k - \sum_{j \in \mathcal{S}} W_{jk} \eta_j \right) &\leq 2 \times ||G_k(\mathbf{W})^\top \beta||_2 \times ||\eta_k - \sum_{j \in \mathcal{S}} W_{jk} \eta_j||_2 \\ &\leq 2 \times ||G_k(\mathbf{W})||_2 \times ||\beta||_2 \times 2\overline{\eta} \\ &\leq 2 \times ||G_k(\mathbf{W})||_2^2 \times ||\beta||_2 \times 2\overline{\eta} \end{aligned}$$

where the first line from Hölder's inequality and the second from Cauchy–Schwarz inequality and the bound for $|\eta_k|$. The final line again adds multiplication by $||G_k(\mathbf{W})||_2$, which can be made greater than 1 (as long as $||G_k(\mathbf{W})||_2 > 0$) by appropriately defining the scale of $g(\mathbf{X})$ and β. This final step is added for simpler interpretation because this bound can be combined together with the bound for the first term and the fifth term.

For the seventh term, we obtain

$$\begin{aligned} 2F_k(\mathbf{W}) \left(\eta_k - \sum_{j \in \mathcal{S}} W_{jk} \eta_j \right) &\leq 2 \times ||F_k(\mathbf{W})||_2 \times ||\eta_k - \sum_{j \in \mathcal{S}} W_{jk} \eta_j||_2 \\ &\leq 2 \times ||F_k(\mathbf{W})||_2 \times 2\overline{\eta} \\ &\leq 4\overline{\eta} \times \rho \times \sum_{j \in \mathcal{S}} W_{jk} ||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2 \\ &\leq 4\overline{\eta} \times \rho \times \sum_{j \in \mathcal{S}} W_{jk} ||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2^2 \end{aligned}$$

where the first line from Hölder's inequality, the second from the bound for $|\eta_k|$, and the third from the bound for $||F_k(\mathbf{W})||_2$ derived above. The final line adds multiplication by $||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2$, which can be made greater than 1 (as long as $||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2 > 0$) by appropriately defining the scale of $g(\mathbf{X})$ and β. This final step is added for simpler interpretation because this bound can be combined together with the bound for the second term.

For the eighth term,

$$\mathbb{E}\left\{ (\sum_{j \in \mathcal{S}} W_{jk} d_j)^2 \right\} = \sum_{j \in \mathcal{S}} W_{jk}^2 \mathbb{E}(d_j^2)$$

$$
\begin{aligned}
&= \sum_{j\in\mathcal{S}} W_{jk}^2 \mathrm{Var}(\widehat{\theta}_j) \\
&\leq \max_{j'\in[N]} \mathrm{Var}(\widehat{\theta}_{j'}) \sum_{j\in\mathcal{S}} W_{jk}^2
\end{aligned}
$$

where $\mathrm{Var}(\widehat{\theta}_j)$ is the variance of the site-specific ATE estimator where site j is a selected study site. The first line follows from $\mathbb{E}(d_j d_j') = 0$ when $j \neq j'$, the second from the definition of variance, and the final line follows from the definition of the maximum. Importantly, $\mathrm{Var}(\widehat{\theta}_j)$ is unknown to researchers at the site-selection stage.

Therefore, taken all together,

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\theta_k - \widehat{\theta}_k^W\right)^2\right\} \\
\leq\ & \lambda_{1k} \times ||G_k(\mathbf{W})||_2^2 + \lambda_{2k} \times \sum_{j\in\mathcal{S}} W_{jk} ||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2^2 + \lambda_{3k} \times \sum_{j\in\mathcal{S}} W_{jk}^2 + 4\overline{\eta}^2
\end{aligned}
$$

where

$$
\begin{aligned}
\lambda_{1k} &\coloneqq ||\beta||_2 \times \left(||\beta||_2 + 2\rho \times \max_{j'\in[N]} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'\in[N]})||_2 + 4\overline{\eta}\right) \\
\lambda_{2k} &\coloneqq \rho^2 \times \frac{\max_{j'} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2}{\min_{j'} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2} + 4\rho\overline{\eta} \\
\lambda_{3k} &\coloneqq \overline{\eta}^2 + \max_{j'\in[N]} \mathrm{Var}(\widehat{\theta}_{j'}).
\end{aligned}
$$

Finally, we take the average of the MSE over non-selected sites.

$$
\begin{aligned}
&\frac{1}{N-N_S} \sum_{k\in\mathcal{R}} \mathbb{E}\left\{\left(\theta_k - \widehat{\theta}_k^W\right)^2\right\} \\
\leq\ & \lambda_1 \times \frac{1}{N-N_S} \sum_{k\in\mathcal{R}} \frac{1}{L_g} ||G_k(\mathbf{W})||_2^2 \\
&+ \lambda_2 \times \frac{1}{N-N_S} \sum_{k\in\mathcal{R}} \sum_{j\in\mathcal{S}} W_{jk} \frac{1}{L_g} ||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2^2 \\
&+ \lambda_3 \times \frac{1}{N-N_S} \sum_{k\in\mathcal{R}} \sum_{j\in\mathcal{S}} W_{jk}^2 \quad + \quad 4\overline{\eta}^2 \\
=\ & \lambda_1 \times \frac{1}{N-N_S} \sum_{k=1}^{N} (1-S_k) \left(\frac{1}{L_g} \sum_{\ell=1}^{L_g} (g_\ell(\mathbf{X}_k) - \sum_{j=1}^{N} S_j W_{jk} g_\ell(\mathbf{X}_j))^2\right) \\
&+ \lambda_2 \times \frac{1}{(N-N_S)} \sum_{j=1}^{N} \sum_{k=1}^{N} W_{jk} S_j (1-S_k) \frac{1}{L_g} \sum_{\ell=1}^{L_g} (g_\ell(\mathbf{X}_j) - g_\ell(\mathbf{X}_k))^2 \\
&+ \lambda_3 \times \frac{1}{N-N_S} \sum_{j=1}^{N} \sum_{k=1}^{N} S_j (1-S_k) W_{jk}^2 \quad + \quad 4\overline{\eta}^2
\end{aligned}
$$

where

$$
\lambda_1 \coloneqq L_g \times ||\beta||_2 \times \left(||\beta||_2 + 2\rho \times \max_{j',k} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2 + 4\overline{\eta}\right)
$$

$$
\begin{aligned}
\lambda_2 &\coloneqq L_g \times \rho^2 \times \max_{k\in[N]} \frac{\max_{j'\in[N]} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2}{\min_{j'\in[N]} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2} + 4\rho\overline{\eta} \\
\lambda_3 &\coloneqq \overline{\eta}^2 + \max_{j'\in[N]} \mathrm{Var}(\widehat{\theta}_{j'}).
\end{aligned}
$$

When we set $C = 4\overline{\eta}^2$, this proves the proposed bound (equation (OA.4)).

This worst-case MSE and analytical expression of $(\lambda_1, \lambda_2, \lambda_3)$ provide important insights. First, when the linearity part $g(\mathbf{X}_k)^\top \beta$ explains a larger amount of across-site heterogeneity, $||\beta||_2$ is larger, which leads to a larger value of λ_1. This will prioritize the first term in the objective function (equation (OA.1)) such that observed site-level variables of non-selected sites are well approximated by those of selected sites. Second, when the underlying model deviates more from a linear model, the residual non-linearity modeled by the Lipschitz function $f(\cdot)$ is more important and ρ is larger, which leads to a larger value of λ_2 (λ_2 includes the quadratic term of ρ, while λ_1 only has the linear term). This will prioritize the second term in the objective function (equation (OA.2)) such that we select sites closer to non-selected sites to avoid excessive reliance on linearity.

Third, when variance of the site-specific ATE in selected sites are large (i.e., $\mathrm{Var}(\widehat{\theta}_j)$ is larger), λ_3 will be larger and the SPS will prioritize the third term the objective function (equation (OA.3)) such that weights are closer to uniform and the downstream weighted average estimator has smaller variance. Finally, when unobserved moderators have larger effects (i.e., $\overline{\eta}$ is larger), λ_3 will be larger (λ_3 includes the quadratic term of $\overline{\eta}$, while λ_1 and λ_2 only have the linear term) and the SPS will prioritize the third term in the objective function (equation (OA.3)) such that site selection and weights estimation depend less on observed site-level variables. □

B.2 Solving SPS Optimization Problem

In this section, we discuss how to solve the SPS optimization problem.

$$
\begin{aligned}
\min_{(\mathbf{S}\in\{0,1\},\ \mathbf{W})} \quad & \lambda_1 \times \frac{1}{N-N_S} \sum_{k=1}^{N} (1-S_k) \left(\frac{1}{L_g} \sum_{\ell=1}^{L_g} \left(g_\ell(\mathbf{X}_k) - \sum_{j=1}^{N} S_j W_{jk} g_\ell(\mathbf{X}_j) \right)^2 \right) \\
& + \lambda_2 \times \frac{1}{(N-N_S)} \sum_{j=1}^{N} \sum_{k=1}^{N} W_{jk} S_j (1-S_k) \frac{1}{L_g} \sum_{\ell=1}^{L_g} (g_\ell(\mathbf{X}_j) - g_\ell(\mathbf{X}_k))^2 \\
& + \lambda_3 \times \frac{1}{N-N_S} \sum_{j=1}^{N} \sum_{k=1}^{N} S_j (1-S_k) W_{jk}^2 \\
\text{such that} \quad & \sum_{k=1}^{N} S_k = N_S, \quad \mathbf{W} \geq 0, \text{and } \sum_j S_j W_{jk} = 1 \ \text{ for all non-selected sites } k \text{ with } S_k = 0.
\end{aligned}
$$

Researchers can also add additional constraints to this problem.

This is a mixed integer programming problem, and we follow techniques in Doudchenko *et al.* (2021) to make the problem quadratic. In particular, we will use the following two auxiliary variables.

$$
Q_{jk} = S_j W_{jk}
$$

$$Z_{k\ell} \quad = \quad (1-S_k)g_\ell(\mathbf{X}_k) - \sum_{j=1}^{N} Q_{jk} g_\ell(\mathbf{X}_j)$$

Using these auxiliary variables, we can rewrite the optimization problem as follows.

$$\begin{aligned}
\min_{(\mathbf{S}\in\{0,1\},\ \mathbf{W},\mathbf{Q},\mathbf{Z})} \quad & \lambda_1 \times \frac{1}{(N-N_S)L_g} \sum_{k=1}^{N}\sum_{\ell=1}^{L_g} Z_{k\ell}^2 && \text{(OA.5)}\\
& + \lambda_2 \times \frac{1}{(N-N_S)} \sum_{j=1}^{N}\sum_{k=1}^{N} Q_{jk} \frac{1}{L_g}\sum_{\ell=1}^{L_g} (g_\ell(\mathbf{X}_j) - g_\ell(\mathbf{X}_k))^2, && \text{(OA.6)}\\
& + \lambda_3 \times \frac{1}{N-N_S} \sum_{j=1}^{N}\sum_{k=1}^{N} Q_{jk}^2 && \text{(OA.7)}\\
\text{such that} \quad & \sum_{k=1}^{N} S_k = N_S, \quad \mathbf{W} \geq 0, \ \sum_{j=1}^{N} Q_{jk} = 1 - S_k, W_{jk} \leq 1 - S_k, && \text{(OA.8)}\\
& Z_{k\ell} = (1-S_k)g_\ell(\mathbf{X}_k) - \sum_{j=1}^{N} Q_{jk} g_\ell(\mathbf{X}_j) && \text{(OA.9)}\\
& 0 \leq Q_{jk} \leq S_j, \text{ and } \ W_{jk} - (1-S_j) \leq Q_{jk} \leq W_{jk} && \text{(OA.10)}
\end{aligned}$$

This is a mixed integer programming problem where the objective function is quadratic and constraints are linear, so any academic and commercial solvers (like CVX) can solve this efficiently.

Proof. We prove this equivalence step by step. We follow techniques in Doudchenko *et al.* (2021). As for the first part of the objective function (equation (OA.5)), we have

$$\begin{aligned}
Z_{k\ell} \quad &= \quad (1-S_k)g_\ell(\mathbf{X}_k) - \sum_{j=1}^{N} Q_{jk} g_\ell(\mathbf{X}_j)\\
&= \quad (1-S_k)g_\ell(\mathbf{X}_k) - \sum_{j=1}^{N} S_j W_{jk} g_\ell(\mathbf{X}_j)\\
&= \quad (1-S_k)g_\ell(\mathbf{X}_k) - (1-S_k)\sum_{j=1}^{N} S_j W_{jk} g_\ell(\mathbf{X}_j)\\
&= \quad (1-S_k)(g_\ell(\mathbf{X}_k) - \sum_{j=1}^{N} S_j W_{jk} g_\ell(\mathbf{X}_j)),
\end{aligned}$$

where the first and second lines follow from definitions, the third from the fact that $W_{jk} = 0$ when $S_k = 1$, and the last line from rearrangement. Therefore,

$$Z_{k\ell}^2 \quad = \quad (1-S_k)(g_\ell(\mathbf{X}_k) - \sum_{j=1}^{N} S_j W_{jk} g_\ell(\mathbf{X}_j))^2,$$

given that $(1-S_k)^2 = (1-S_k)$.

As for the second part of the objective function (equation (OA.6)), we have

$$W_{jk}S_j(1-S_k) = S_jW_{jk} = Q_{jk}.$$

because $W_{jk} = 0$ when $S_k = 1$ and we use the definition of Q_{jk}

As for the third part of the objective function (equation (OA.7)), we have

$$S_j(1-S_k)W_{jk}^2 = S_jW_{jk}^2 = Q_{jk}^2,$$

because $W_{jk} = 0$ when $S_k = 1$ and $S_j^2 = S_j$.

As for constraints, the first two constraints are the same as before. $\sum_{j=1}^{N} Q_{jk} = 1 - S_k$ is equivalent to $\sum_j S_jW_{jk} = 1$ for all non-selected sites k with $S_k = 0$. $W_{jk} \leq 1 - S_k$ makes sure that $W_{jk} = 0$ when $S_k = 1$. Equation (OA.9) defines $Z_{k\ell}$. Equation (OA.10) defines Q_{jk} only using linear rules. When $S_j = 1$, equation (OA.10) implies that $0 \leq Q_{jk} \leq 1$ and $W_{jk} \leq Q_{jk} \leq W_{jk}$, and thus, $Q_{jk} = W_{jk}$. Instead, when $S_j = 0$, equation (OA.10) implies that $0 \leq Q_{jk} \leq 0$ and $W_{jk} - 1 \leq Q_{jk} \leq W_{jk}$, and thus, $Q_{jk} = 0$. When we combine both cases, equation (OA.10) is equivalent to $Q_{jk} = S_jW_{jk}$. This completes the proof. □

B.3 SPS Estimators

B.3.1 Weighted Average Estimator

We show that the SPS estimator is a weighted average of the site-specific ATE estimators in selected sites.

Proof. First, we have

$$\begin{aligned}
\widehat{\theta}_{AS} &\coloneqq \frac{1}{N}\Big(\sum_{j\in\mathcal{S}}\widehat{\theta}_j + \sum_{k\in\mathcal{R}}\widehat{\theta}_k^W\Big) \\
&= \frac{1}{N}\Big(\sum_{j\in\mathcal{S}}\widehat{\theta}_j + \sum_{k\in\mathcal{R}}\sum_{j\in\mathcal{S}}\widehat{W}_{jk}\widehat{\theta}_j\Big) \\
&= \frac{1}{N}\Big(\sum_{j\in\mathcal{S}}\widehat{\theta}_j + \sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{R}}\widehat{W}_{jk}\widehat{\theta}_j\Big) \\
&= \frac{1}{N}\sum_{j\in\mathcal{S}}\Big(1 + \sum_{k\in\mathcal{R}}\widehat{W}_{jk}\Big)\widehat{\theta}_j,
\end{aligned}$$

where the first and second lines follow from the definition of the SPS estimators, the third equality follows from exchange of two summations, and the last equality from rearrangement of terms.

Therefore, when we define $\widetilde{W}_j = (1 + \sum_{k\in\mathcal{R}}\widehat{W}_{jk})/N$, we have

$$\widehat{\theta}_{AS} = \sum_{j\in\mathcal{S}}\widetilde{W}_j\widehat{\theta}_j.$$

Finally, we need to check $\sum_{j\in\mathcal{S}}\widetilde{W}_j = 1$.

$$\sum_{j\in\mathcal{S}}\widetilde{W}_j = \frac{1}{N}\sum_{j\in\mathcal{S}}\Big(1 + \sum_{k\in\mathcal{R}}\widehat{W}_{jk}\Big)$$

$$
\begin{aligned}
&= \frac{1}{N}(N_S + \sum_{j\in\mathcal{S}}\sum_{k\in\mathcal{R}} \widehat{W}_{jk}) \\
&= \frac{1}{N}(N_S + \sum_{k\in\mathcal{R}}\sum_{j\in\mathcal{S}} \widehat{W}_{jk}) \\
&= \frac{1}{N}(N_S + N - N_S) \\
&= 1
\end{aligned}
$$

where the first line follows from the definition of $\widetilde{W}_j$, the second from distribution of the sum $\sum_{j\in\mathcal{S}}$, the third from exchange of two summations, the fourth line from the SPS constraint $\sum_{j\in\mathcal{S}} \widehat{W}_{jk} = 1$, and the final step follows from the fact that the number of selected sites is N_S and the number of non-selected sites is $N - N_S$. This completes the proof. □

B.3.2 Variance Estimation

Now we consider variance estimation for the averarge-site ATE estimator. We consider both site-level and unit-level error terms. In particular, without loss of generality, we define

$$
\widehat{\theta}_k = \theta_k + \delta_k + \epsilon_k
$$

where $\widehat{\theta}_k$ is an estimate of the site-specific ATE at site k, θ_k is a constant parameter that represents the true site-specific ATE in site k, δ_k captures a site-level error term, and, ϵ_k captures the within-site error term, which is the within-site average of unit-level error terms. We will analyze δ_k and ϵ_k as random variables.

Importantly, we don't assume θ_k comes from some unknown super-population. θ_k is a fixed constant parameter that represents the true site-specific ATE in site k. ϵ_k is the within-site error term, which is the within-site average of unit-level error terms. Thus, importantly, ϵ_k decreases as sample size within site k increases. When an unbiased estimator is used in site k, $\mathbb{E}(\epsilon_k) = 0$. Because causal studies in each site use independent sets of data, ϵ is independent across sites without loss of generality.

δ_k captures the non-systematic site-level error term that does not vanish even when sample size at site k is infinite. For example, this captures the weather of days when a study is conducted in site k, and random variations of treatment implementation. Even if a study in site k has infinite sample size, an estimate of the site-specific ATE will not be exactly the same if we hypothetically re-run a study many times due to such random site-level variations. δ_k caputures this inherent site-level non-systematic random variation, whereas systematic heterogeneity across sites is captured by θ_k. Thus, without loss of generality, $\mathbb{E}(\delta_k) = 0$. We assume site-level error term δ is independent across sites.

Given this basic setup, we can write the variance of the SPS estimator as follows.

$$
\begin{aligned}
\mathrm{Var}(\widehat{\theta}_{AS}) &= \sum_{j\in\mathcal{S}} \widetilde{W}_j^2 \mathrm{Var}(\widehat{\theta}_j) \\
&= \sum_{j\in\mathcal{S}} \widetilde{W}_j^2 (\sigma_j^2 + \tau^2)
\end{aligned}
$$

where $\sigma_j^2 = \text{Var}(\epsilon_j)$, which is the within-site variance of the site-specific ATE estimator, and $\tau^2 = \text{Var}(\delta)$, which is the across-site variance. Recall that $\mathcal{S}$ is a set of selected study sites. $\text{Var}(\cdot)$ is defined as the variance over random variables ϵ and δ. Importantly, as in typical experimental analysis, we consider randomness conditional on the design stage (i.e., observed covariates $\mathbf{X}$) and thus, $\widetilde{W}$ are treated as constants.

We can easily obtain an estimate of the within-site variance σ_k^2 for site k using an estimated variance of the site-specific ATE estimator in site k.

We now turn to estimation of the across-site variance τ^2. We will show below the following variance estimator is a conservative variance estimator, i.e., $\mathbb{E}(\widehat{\tau}^2) \geq \tau^2$.

$$\widehat{\tau}^2 \coloneqq \frac{\sum_{k\in\mathcal{S}} \widehat{e}_k^2 - (\sum_{k\in\mathcal{S}} \widehat{\sigma}_k^2 + \sum_{k\in\mathcal{S}} \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}^2 \widehat{\sigma}_j^2)}{\sum_{k\in\mathcal{S}}(1 + \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}^2)},$$

where

$$\widehat{e}_k \;\coloneqq\; \widehat{\theta}_k - \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}\widehat{\theta}_j,$$

$\widehat{\sigma}_k^2$ is an (asymptotically) unbiased estimate of the within-site variance σ_k^2. $\mathcal{S}_k$ is a set of selected sites after removing site k. $\overline{W}_{jk}$ is the SPS weight we estimate to approximate site k only using sites in $\mathcal{S}_k$. Importantly, this variance estimator does not assume that our SPS estimator is unbiased. This variance estimator is valid even when the SPS estimator is biased.

Taken together,

$$\widehat{\text{Var}}(\widehat{\theta}_{AS}) \;=\; \sum_{j\in\mathcal{S}} \widetilde{W}_j^2(\widehat{\sigma}_j^2 + \widehat{\tau}^2).$$

Proof. We now prove the property of $\widehat{\tau}^2$. We start with the decomposition of $\widehat{e}_k$.

$$\begin{aligned}
\widehat{e}_k \;&\coloneqq\; \widehat{\theta}_k - \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}\widehat{\theta}_j \\
&=\; (\theta_k - \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}\theta_j) + (\delta_k - \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}\delta_j) + (\epsilon_k - \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}\epsilon_j),
\end{aligned}$$

where $(\theta_k - \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}\theta_j)$ is a fixed constant and does not contain randomness. We will use b_k to denote this unknown bias term. Therefore,

$$\begin{aligned}
\mathbb{E}(\widehat{e}_k^2) \;&=\; \text{Var}(\widehat{e}_k) + \mathbb{E}(\widehat{e}_k)^2 \\
&=\; \left(\text{Var}(\delta_k) + \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}^2 \text{Var}(\delta_j)\right) + \left(\text{Var}(\epsilon_k) + \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}^2 \text{Var}(\epsilon_j)\right) + b_k^2 \\
&=\; (1 + \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}^2)\tau^2 + \sigma_k^2 + \sum_{j\in\mathcal{S}_k} \overline{W}_{jk}^2 \sigma_j^2 + b_k^2
\end{aligned}$$

where the first equality follows from the definition of variance, the second from the decomposition above, and the third from definitions of τ^2 and σ_k^2.

Averaging over sites, we obtain

$$\frac{1}{N_S}\sum_{k\in\mathcal{S}}\mathbb{E}(\widehat{e}_k^2) \;=\; \tau^2 \times \frac{1}{N_S}\sum_{k\in\mathcal{S}}(1+\sum_{j\in\mathcal{S}_k}\overline{W}_{jk}^2) + \frac{1}{N_S}\sum_{k\in\mathcal{S}}\sigma_k^2 + \frac{1}{N_S}\sum_{k\in\mathcal{S}}\sum_{j\in\mathcal{S}_k}\overline{W}_{jk}^2\sigma_j^2 + \frac{1}{N_S}\sum_{k\in\mathcal{S}} b_k^2.$$

Rearranging the term, we have

$$\tau^2 = \frac{\sum_{k\in\mathcal{S}}\mathbb{E}(\widehat{e}_k^2) - (\sum_{k\in\mathcal{S}}\sigma_k^2 + \sum_{k\in\mathcal{S}}\sum_{j\in\mathcal{S}_k}\overline{W}_{jk}^2\sigma_j^2) - \sum_{k\in\mathcal{S}} b_k^2}{\sum_{k\in\mathcal{S}}(1+\sum_{j\in\mathcal{S}_k}\overline{W}_{jk}^2)}.$$

We will replace $\mathbb{E}(\widehat{e}_k^2)$ with an unbiased estimator $\widehat{e}_k^2$, and we will replace σ_k^2, and σ_j^2 with (asymptotically) unbiased estimators $\widehat{\sigma}_k^2$, and $\widehat{\sigma}_j^2$. Importantly, $\sum_{k\in\mathcal{S}} b_k^2$ is unknown and unestimable, but we know it is equal to or greater than zero $\sum_{k\in\mathcal{S}} b_k^2 \geq 0$.

Therefore,

$$\begin{aligned}
\mathbb{E}(\widehat{\tau}^2) &= \frac{\sum_{k\in\mathcal{S}}\mathbb{E}(\widehat{e}_k^2) - (\sum_{k\in\mathcal{S}}\mathbb{E}(\widehat{\sigma}_k^2) + \sum_{k\in\mathcal{S}}\sum_{j\in\mathcal{S}_k}\overline{W}_{jk}^2\mathbb{E}(\widehat{\sigma}_j^2))}{\sum_{k\in\mathcal{S}}(1+\sum_{j\in\mathcal{S}_k}\overline{W}_{jk}^2)} \\
&= \frac{\sum_{k\in\mathcal{S}}\mathbb{E}(\widehat{e}_k^2) - (\sum_{k\in\mathcal{S}}\sigma_k^2 + \sum_{k\in\mathcal{S}}\sum_{j\in\mathcal{S}_k}\overline{W}_{jk}^2\sigma_j^2)}{\sum_{k\in\mathcal{S}}(1+\sum_{j\in\mathcal{S}_k}\overline{W}_{jk}^2)} \\
&\geq \frac{\sum_{k\in\mathcal{S}}\mathbb{E}(\widehat{e}_k^2) - (\sum_{k\in\mathcal{S}}\sigma_k^2 + \sum_{k\in\mathcal{S}}\sum_{j\in\mathcal{S}_k}\overline{W}_{jk}^2\sigma_j^2) - \sum_{k\in\mathcal{S}} b_k^2}{\sum_{k\in\mathcal{S}}(1+\sum_{j\in\mathcal{S}_k}\overline{W}_{jk}^2)} \\
&= \tau^2.
\end{aligned}$$

Importantly, this variance estimator does not assume the SPS estimator is unbiased. When the SPS estimator is indeed unbiased (i.e., $b_k = 0$), our variance estimator is also unbiased. However, even when the SPS estimator is biased, our variance estimator is guaranteed to be conservative. □

B.3.3 Inference

To make inference, we have to handle site-level error terms and within-site error terms. First, with large-sample approximation, we can use the central limit theorem to prove that

$$\widehat{\theta}_k \sim \mathcal{N}(\theta_k^*, \sigma_k^2),$$

where $\theta_k^* = \theta_k + \delta_k$.

However, as for site-level randomness, because the number of study sites is often small in political science and other social science fields, unfortunately, we cannot use large sample approximation. Instead, we follow the standard literature of meta-analysis (DerSimonian and Laird, 1986) and make a distributional assumption. In particular, we assume

$$\theta_k^* \sim \mathcal{N}(\theta_k, \tau^2).$$

Importantly, this is distinct from a random-effect meta-analysis model in a fundamental way: the mean θ_k is a constant site-specific ATE, and we don't assume any global superpopulation

of sites. This is also different from a fixed-effect meta-analysis model in that we explicitly take into account across-site heterogeneity.

Given this normal assumption, we obtain

$$\widehat{\theta}_k \sim \mathcal{N}(\theta_k, \sigma_k^2 + \tau^2),$$

and thus,

$$\widehat{\theta}_{AS} \sim \mathcal{N}(\sum_{j \in S} \widetilde{W}_j \theta_j, \mathrm{Var}(\widehat{\theta}_{AS})).$$

In practice, we use the proposed conservative variance estimator $\widehat{\mathrm{Var}}(\widehat{\theta}_{AS})$ to obtain conservative confidence intervals and p-values.