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

Naoki Egami[†] Diana Da In Lee[‡]

First Version: July 3, 2023
This Version: August 25, 2024

Abstract

Multi-site/context studies have become popular strategies to address the most common and challenging external validity concerns about contexts. Under such studies, scholars conduct causal studies in each site and evaluate whether findings generalize across sites. Despite the potential, there has been little guidance on the fundamental research design question—how should we select sites for external validity? Existing approaches have challenges: random sampling of sites is often infeasible, while the current practice of purposive sampling is suboptimal without statistical guarantees. We propose *synthetic purposive sampling* (SPS), which optimally selects diverse sites for external validity. SPS combines ideas of purposive sampling and the synthetic control method—it selects diverse sites such that non-selected sites are well approximated by the weighted average of the selected sites. We illustrate its general applicability using both experimental and observational studies. Overall, this paper offers a new statistical foundation to design multi-site studies for external validity.

[*] 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. We would like to thank Brandon de la Cuesta, Michael Denly, Michael Findley, Don Green, Jens Hainmueller, Melody Huang, Kosuke Imai, Eunji Kim, Gary King, Shiro Kuriwaki, Adeline Lo, Ian Lundberg, John Marshall, Kevin Munger, Jonathan Renshon, Cyrus Samii, Fredrik Sävje, Tara Slough, Elizabeth Tipton, Scott Tyson, Chagai Weiss, Anna Wilke, Teppei Yamamoto, and Yang-Yang Zhou, 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. Egami acknowledges financial support from the National Science Foundation (SES–2318659).

[†] 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 science has 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 revolves around *external validity* (Shadish, Cook and Campbell, 2002; Egami and Hartman, 2023). While the concept of external validity is multi-dimensional, the essential question to social scientists involves contexts: how can researchers generalize causal findings across different contexts? For example, do experimental findings about voter information campaigns in Kenya generalize to other countries in Africa? What do causal findings about partisan bias in several US cities teach us about partisan bias more generally in the US? These are some of the most common and yet most challenging external validity concerns social scientists face in practice across disciplines.

A promising strategy to address this question is a multi-site/multi-context causal study where researchers conduct experimental or observational studies in multiple contexts to compare and aggregate findings across contexts.[1] 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, Cook and Campbell, 2002; Blair and McClendon, 2020).

Recognizing the importance of external validity, an increasing number of scholars deploy 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 has increased steadily since then. One popular type is a multi-country survey experiment that tests how causal findings vary across countries (e.g., Tomz and Weeks, 2013; Valentino et al., 2019; Arechar et al., 2023), and Bassan-Nygate et al. (2023) is a recent prominent example testing the external validity of well-known IR experimental findings. To understand the growing trend of multi-site causal studies in a broader picture, we

[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.

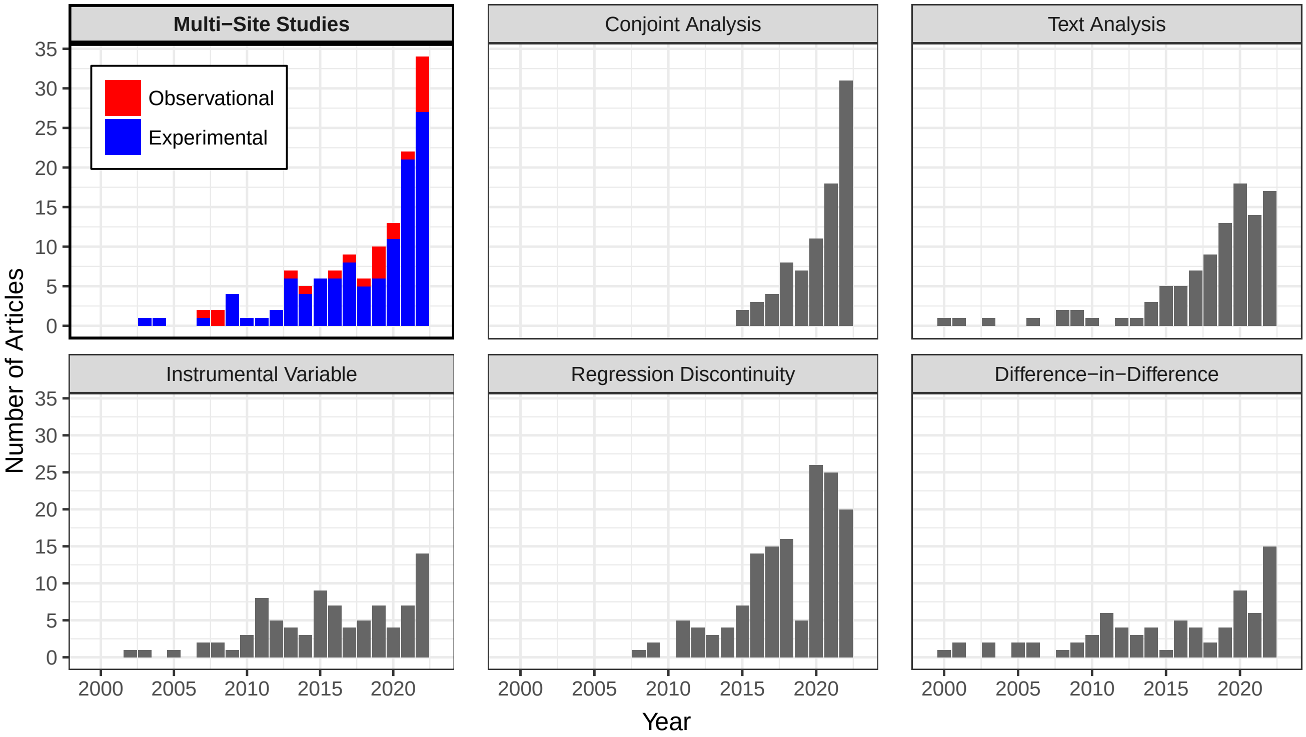


Figure 1: **Increasingly Popular Multi-Site Causal Studies.**
Note: In the top left panel, blue (red) bars represent multi-site experimental (observational) studies. The plots are based on a review of articles published in the top 10 political science journals from 2000 to 2022. See Appendix D for more details on the literature review process.

also counted the number of papers using other established methods in the same top 10 political science journals. We find that the recent increase in multi-site causal studies is comparable to that of conjoint analysis, and the number of multi-site studies has already exceeded the number of papers using other well-known methods, such as text analysis and instrumental variables, in recent years. 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 across contexts. In addition, multi-site causal studies have been strongly supported by initiatives like Metaketa by EGAP (e.g., Blair et al., 2021). Overall, Figure 1 shows an exciting pattern that multi-site causal studies have become widely used not only by large-scale coordinated efforts like Metaketa but also by many individual researchers. Many papers on external validity also endorse multi-context studies explicitly or implicitly (e.g., McDermott, 2011; Findley, Kikuta and Denly, 2020; Egami and Hartman, 2023; Slough and Tyson, 2022; Wilke and Samii, 2023).

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? This question of site selection is essential because causal findings in the social sciences can substantially vary across study sites that differ in political, economic, demographic, and other characteristics. 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.

Broadly speaking, there are two classes of existing strategies: random sampling and purposive sampling. First, if feasible, random sampling of sites is a powerful approach to make generalizable causal claims from multi-site studies. Randomly selected sites are representative, and researchers are protected from known and unknown systematic bias in site selection. However, unfortunately, in practice, random sampling from all the sites of theoretical interest is often infeasible because of logistical and ethical reasons. Indeed, our literature review found only two studies that use random sampling (less than 2 % of all multi-site studies we reviewed).

Given the difficulty of random sampling, researchers often rely on purposive sampling (Shadish, Cook and Campbell, 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 the chosen study sites cover heterogeneous contextual factors (about 80% of multi-site studies justify their site selection in this manner). For example, when studying attitudes toward immigrants using survey experiments (e.g., Naumann, F. Stoetzer and Pietrantuono, 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 theoretically relevant factors exist (our literature review finds that the average number of covariates researchers diversify is 2.17). Second, the process of purposive sampling is often not transparent

	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 incorporating ideas from the synthetic control method (SCM).

or 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 incorporating ideas from the synthetic control method (Abadie, Diamond and Hainmueller, 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. We also propose a corresponding SPS estimator to aggregate causal estimates from selected study sites.

The proposed SPS overcomes the shortcomings of existing methods (see Table 1). First, it is a flexible approach. SPS can accommodate logistical and practical constraints. For example, where researchers can run experiments is often constrained by non-theory-driven reasons, such as funding opportunities, availability of collaborators, and knowledge of the local context and language. Rather than hiding such real constraints social scientists face, SPS will select the optimal set of study sites within such user-specified constraints. Researchers can also incorporate theoretical and domain knowledge by combining classical purposive sampling and SPS. For example, scholars can first select one or two sites based on qualitative knowledge (e.g., because the sites are typical of a given substantive theory of interest or because they provide a hard test) and then use SPS for selecting the remaining sites to complement and improve the diversity of study sites. 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 theoretically relevant 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. 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 a statistical foundation, whereas existing methods offer only one of these features.

Our algorithm is general and is designed to help researchers systematically select study sites in a wide range of applications. For example, researchers can use SPS to select diverse countries for multi-country causal studies. SPS can also be used to select sites within a country when selecting cities, districts, states, schools, and so on. To illustrate the general applicability, we demonstrate the use of SPS with five empirical applications across different subfields, ranging from experimental to observational studies and from studies conducted in a few sites to studies conducted in more than 10 sites (see Table 2). We also offer a companion `R` package `spsR`, which can implement all the methods described in this paper (see Appendix B for an introduction to the package with code examples).

Overall, this paper offers a new statistical foundation to design multi-site studies for external validity. We take a *prospective* approach to explicitly design multi-site studies for external validity upfront *before* data collection. While it is currently common to think about external validity only at the final stage of studies *after* data collection (e.g., when writing up papers), such post hoc adjustment requires strong and often untenable assumptions, especially when external validity concerns are about contexts. SPS allows researchers to explicitly address external validity concerns about contexts upfront through their research design.

In the next section, we begin with several motivating applications. After reviewing the challenges of existing methods (Section 3), we introduce SPS (Section 4) and discuss how to aggregate causal evidence from multiple sites (Section 5). We then reanalyze the empirical applications introduced earlier (Section 6). We offer practical recommendations and clarify precautions in Section 7. In Section 8, we discuss connections to other important literature, such as case selection in qualitative studies.

Related Literature

This paper builds on several lines of work. First, we contribute to the growing literature on external validity (e.g., Shadish, Cook and Campbell, 2002; Tipton, 2013; Allcott, 2015; Bareinboim and Pearl, 2016; Coppock, Leeper and Mullinix, 2018; Meager, 2019; Munger, 2019; Blair and McClendon, 2020; Findley, Kikuta and Denly, 2020; Vivalt, 2020; Egami and Hartman, 2021; Miratrix, Weiss and Henderson, 2021; Chassang and Kapon, 2022; Devaux and Egami, 2022; Slough and Tyson, 2022; Bassan-Nygate et al., 2023; Egami and Hartman, 2023; Findley, Denly and Kikuta, 2023; Wilke and Samii, 2023). 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 discuss the potential use of SPS for other dimensions of external validity 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; Gechter et al., 2023). 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

Paper	**Topic**	**Type of Causal Studies**	**Type of Sites**	**# of Sites**
Gift & Gift (2015)	Partisan Bias in Hiring in the US	Audit Experiments	Counties in the US	2
Naumann et al. (2018)	Attitudes towards Immigrants	Survey Experiments	European Countries	15
Blair et al. (2021)	Community Policing (Metaketa)	Field Experiments	Countries in the Global South	6
Lupu & Wallace (2019)	International Human Rights Law	Survey Experiments	Countries across Continents	3
Bisbee et al. (2017)	Effect of Fertility on Labor Supply	Observational Studies	Countries across Continents	45

Table 2: **A Wide Range of Empirical Applications Analyzed in this Paper.**
Note: Given space constraints, we offer the first three applications in the main text and the last two applications in Appendix A.

control method, which was also developed for the small sample regime.

Finally, our paper builds on the literature on the synthetic control method (Abadie, Diamond and Hainmueller, 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 leads 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 Motivating Empirical Applications

To demonstrate how our approach can improve a wide range of multi-site studies, we use a diverse set of empirical applications (see Table 2). Examples cover different types of causal studies (field and survey experiments as well as observational studies), numbers of sites (from small to moderate and large), types of sites (countries across continents and counties within the US), and subfields (American politics, comparative politics, and international relations).

In this section, we briefly describe three empirical applications, which serve as illustrative examples throughout the paper. Given that a vast majority of existing multi-site studies are experimental (see Figure 1), we use multi-site experiments as the main examples. We offer two additional empirical applications, including an observational study, in Appendix A.

2.1 Field Experiments in US Counties: Partisan Bias in Hiring

In the age of polarization, scholars have found that partisanship influences not only political but also economic and social domains in the US. In an influential study, Gift and Gift (2015) conducted field audit experiments in two US counties—one highly conservative and one highly liberal—to examine whether partisan signals affect hiring. In particular, within each county, they sent out politically branded resumes that randomly included liberal, conservative, or no partisan signals. The authors found that job candidates with out-partisan affiliations are less likely to obtain a callback than candidates without any partisan affiliation. We will use this work to illustrate how researchers can apply SPS to systematically select a small number of diverse study sites for external validity.

2.2 Survey Experiments in Europe: Attitudes toward Immigrants

A long-standing question in the immigration literature asks whether natives prefer high-skilled migrants to low-skilled migrants. Naumann, F. Stoetzer and Pietrantuono (2018) tackled this question by running survey experiments in 15 European countries that differ in the size of immigrant population, GDP, unemployment rates, and so on. In each country, they conducted a survey experiment where they randomly changed the skill level of hypothetical immigrant groups (high- or low-skilled) and asked respondents to show the support level for a given immigrant group. They found that, in all 15 countries, respondents preferred high-skilled immigrants to low-skilled immigrants, while there are substantial variations in effect size across countries.

This study will serve as an example of how to use the proposed method in increasingly popular, multi-country survey experiments. Such multi-country survey experiments are likely to continue growing as popular online platforms can recruit survey respondents across the world.[2]

2.3 Metaketa in the Global South: Community Policing

A multi-site field experiment has become famous in political science partly due to a large-scale collective effort by EGAP's Metaketa initiative. The most recent Metaketa project by Blair et al. (2021) examines whether community policing can build citizen trust in police and reduce

[2] Lucid offers surveys in more than 130 countries, and YouGov covers more than 70 markets.

crime by conducting coordinated field experiments in six countries in the Global South (Brazil, Colombia, Liberia, Pakistan, Philippines, and Uganda). They found that the community policing intervention did not improve citizen-police relationships or reduce crime. We will use this as an example to demonstrate how the proposed method can help researchers systematically select diverse sites by explicitly accommodating logistical and ethical constraints.

3 Existing Methods and Their Methodological Challenges

3.1 Random Sampling

Random sampling of sites is one of the most powerful strategies for external validity (Shadish, Cook and Campbell, 2002; Fearon and Laitin, 2008). 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 biases in site selection.

Unfortunately, random sampling is often infeasible in social science applications due to logistical and ethical constraints (see also Findley, Kikuta and Denly, 2020). For example, scholars might consider conducting field experiments related to elections in Wisconsin. Yet, it might be ethically and logistically impossible to do so, given that it is a battleground state. Indeed, we find only two studies that use random sampling of sites in our literature review of multi-site studies.

Another challenge of random sampling is that it might be ineffective when the number of study sites is 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 D). When the number of study sites is small, random sampling can be ineffective because fundamental statistical theorems (e.g., the central limit theorem) are not applicable to a sample size that is too small.

We 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, where scientists often have a relatively large number of schools as sites.

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 non-probability sampling techniques that select sites with "theoretical purposes." It has a long history in the research design literature (Shadish, Cook and Campbell, 2002) and has a wide range of well-developed variants, such as typical, extreme, and most similar selections (Seawright and Gerring, 2008).

In practice, the most popular version is to select diverse sites such that the chosen study sites cover a wide range of values in each site-level variable relevant to a substantive theory of interest. For example, in Gift and Gift (2015), the authors selected two US counties that differ in partisanship, one highly conservative and one highly liberal. Naumann, F. Stoetzer and Pietrantuono (2018) examined attitudes toward immigrants using survey experiments in diverse countries that differ in sizes of immigrant populations, GDP, unemployment rates, and so on. In our literature review of the top 10 political science journals, we find that about 80% of multi-site studies 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 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 local partner who can help her run high-quality experiments.

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 relevant factors matter (in our literature review, we find that the average number of covariates researchers diversify is 2.17). Second, the process of purposive sampling is often not transparent or 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, the current practice of purposive sampling can be characterized as "stratified convenience sampling" (Olsen et al., 2013), i.e., researchers carefully consider one or two contextual factors to stratify, but they ultimately opt for the most convenient sites after stratification.

4 Synthetic Purposive Sampling

In this section, we propose a general site selection method for external validity, which we call *synthetic purposive sampling* (SPS). SPS improves upon conventional purposive sampling by combining ideas from the synthetic control method (Abadie, Diamond and Hainmueller, 2010).

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

4.1 Framework for Site Selection

Before introducing SPS, we begin by developing a framework for site selection. We first define N potential sites of interest as the target population of sites, which is the target against which the external validity of a given substantive theory is evaluated. Specifying the target 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, Kikuta and Denly, 2020; Egami and Hartman, 2023). For example, Naumann, F. Stoetzer and Pietrantuono (2018) are interested in countries in Europe where immigration is a salient political issue. We also provide detailed practical guidance in Section 7.1.

In most scenarios, researchers cannot extensively study all N sites of potential interest. Among them, researchers 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 our general methodology also accommodates observational studies (see an application of an observational multi-site study in Appendix A). For example, in our reanalysis of Naumann, F. Stoetzer and Pietrantuono (2018) (Section 6.1), 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 the same treatment and outcome variables to capture the same underlying theoretical concepts in each site (see Slough and Tyson, 2022; Wilke and Samii, 2023).

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 in 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 the 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

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

which represents the average of the ATEs across all N sites of interest, which also includes sites that we did not select. 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 (see, e.g., Gerber and Green, 2012; Blair et al., 2021; Bassan-Nygate et al., 2023).

Researchers are also often interested in testing the implications of a theoretical mechanism by estimating causal effects separately for different subgroups of sites. We can define the *Subgroup Average-Site ATE* (also known as the conditional average-site ATE) as

$$\theta^{g}_{sub} := \frac{1}{N_g} \sum_{k:G_k = g} \theta_k, \tag{2}$$

which represents the average of the ATEs among a subgroup of sites with variable $G_k = g$ where N_g is the number of sites with $G_k = g$. For example, in Naumann, F. Stoetzer and Pietrantuono (2018), researchers might be interested in testing whether the extent to which natives prefer high-skilled immigrants to low-skilled immigrants is stronger in countries with higher fiscal exposure to migration (i.e., the net burden of migration on public finances is higher) than in countries with lower fiscal exposure (see also Valentino et al., 2019). To examine the implications of this theory, researchers can estimate the subgroup average-site ATEs separately for countries with high and low fiscal exposure (i.e., variable G is fiscal exposure and $g \in \{\text{high, low}\}$). By comparing these subgroup average-site ATEs, researchers can systematically explore the across-site heterogeneity of the ATE. As in subgroup analyses and conditional ATE analyses that are standard in single-site experiments, it is important to note that while each of the subgroup average-site ATE is a causal effect, the difference between them is descriptive.[3]

If researchers can randomly sample study sites, it is straightforward to unbiasedly estimate θ_{AS} and θ_{sub}^{g} using the average or the subgroup 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 these causal quantities of interest.

Note that even if researchers are not specifically interested in estimating the (subgroup) average-site ATEs, 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 7.

4.2 The Proposed Methodology

We now introduce *synthetic purposive sampling* (SPS). Like purposive sampling, SPS selects diverse sites. But unlike the current practice of purposive sampling, we design site selection by explicitly taking into account downstream analyses, i.e., how to use selected sites for generalization. In particular, we use weighted average estimators as in the synthetic control method—we 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. Second, it is also a stable estimator that works well with small sample sizes. Finally, it is a familiar estimator to social scientists as most meta-analysis estimators are also weighted average estimators, even though

[3] Note that we discuss connections between our method and meta-regression in Section 7.

how we construct weights is distinct from meta-analysis estimators.

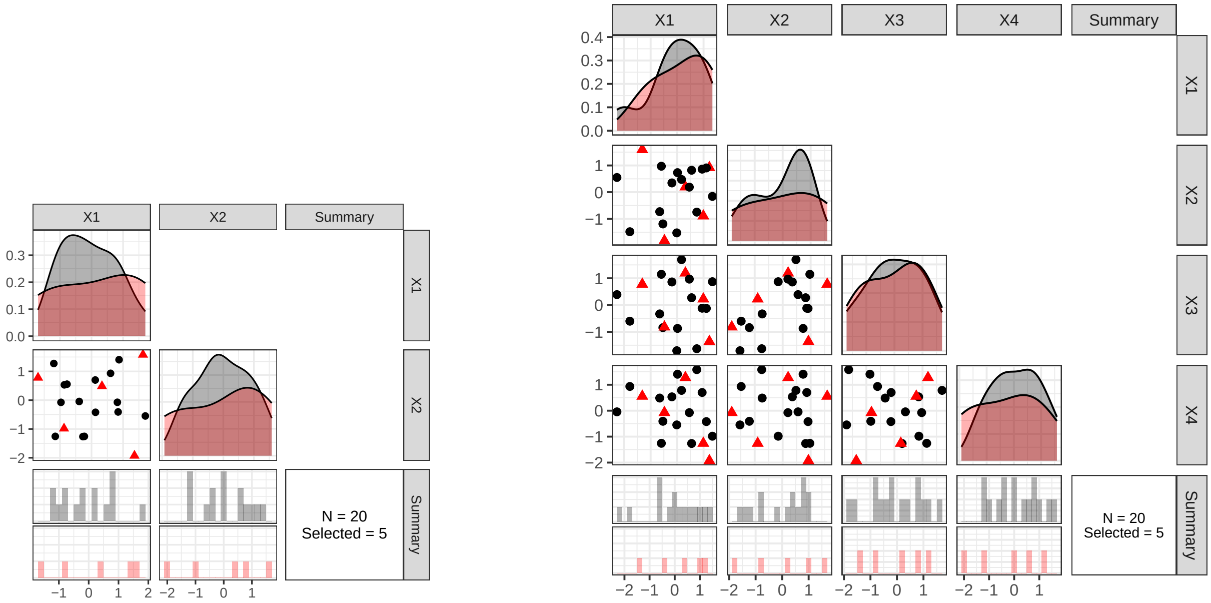
By combining these ideas, SPS will select diverse sites such that non-selected sites can be well approximated by the weighted average of the selected sites. By doing so, even without random sampling, we can make the weighted average of selected sites representative of a population of N sites, including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

More concretely, like existing purposive sampling approaches, the first step of SPS is to choose site-level variables $\mathbf{X}_k = (X_{k1}, X_{k2}, \dots, X_{kL})$ that users want to diversify across sites where L is the number of site-level variables. In particular, researchers should include contextual variables that are theoretically expected to explain differences in the ATEs across sites. This choice needs to be based on the theoretical and domain knowledge of a given application (Findley, Kikuta and Denly, 2020; Bassan-Nygate et al., 2023). For example, to run multi-country survey experiments on attitudes toward immigrants, researchers might diversify the size of immigrant populations, GDP, and unemployment rates, which are key contextual factors discussed by theories in the immigration literature. In a study of partisan bias in hiring by Gift and Gift (2015), the original authors wanted to diversify the unemployment rate because hiring companies in counties with higher unemployment rates are likely to have a larger pool of job candidates such that they might have more rooms for using partisanship to distinguish applicants who are otherwise similar in qualifications.

This step of choosing theoretically relevant site-level variables is what researchers are already doing when diversifying one or two variables by hand in the conventional purposive sampling approach. With SPS, researchers can incorporate any number of theoretically relevant covariates that are expected to moderate the ATEs across sites.[4]

We emphasize that SPS is *not* a substitute for theoretical arguments of the underlying mechanism and domain knowledge essential in site selection. Rather, researchers must use the theoretical and domain knowledge of a given application to choose site-level variables they diversify via SPS. The method we offer *augments* theoretical discussion by systematically and optimally diversifying variables chosen based on domain knowledge in each application.

[4] There are a large number of site-level data sets available across different geographic units, e.g., country-, state-, city-, and district-level data (see Appendix C).



(a) (b)

Figure 2: **Illustration with Simple Simulated Data.**
Note: Panels (a) and (b) consider two and four variables, respectively. Scatter plots show values of site-level variables $\mathbf{X}$ of selected sites (red triangles) and non-selected sites (black circles). Density plots on the diagonal and histograms in the last rows compare the distribution of site-level variables between selected (red) and non-selected (black) sites.

Given the choice of theoretically relevant site-level variables, 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 observed covariates chosen by users. We discuss unobserved moderators and introduce a procedure similar to cross-validation that can empirically assess the potential influence of unobserved site-level variables in Section 5.1, as it is clearer to discuss them after introducing the SPS estimator there.

4.2.1 Illustration

Here, we illustrate SPS with a simple simulation study where researchers choose five sites from the population of 20 potential sites (see Figure 2). When researchers have only two variables to consider (Figure 2-(a)), it is a relatively easy task to choose five sites. If one can select sites close to the center and four corners in the scatter plot, other 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. In this simple example, SPS selected sites as such (selected sites are represented by red triangles 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 Figure 2-(a) show the marginal distributions of each variable.

However, in practice, researchers often want to consider many theoretically relevant variables $\mathbf{X}$ that are predictive of across-site heterogeneity of causal effects. As shown in Figure 2-(b), even when they have only four variables, they have to simultaneously consider six two-dimensional figures. This task becomes 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 consider many variables simultaneously and choose diverse sites. Figure 2-(b) shows that SPS indeed selects diverse sites such that many non-selected sites can be well approximated in all dimensions. As a result, selected sites cover a wide range of values in all four contextual factors (see the last row of Figure 2-(b)).

4.2.2 Optimization Problem behind SPS

To formally introduce SPS, we require some notation. 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 the weight we assign to selected site j when predicting the ATE of non-selected site k. Then, we can define an imbalance measure $B_{k\ell}$ for non-selected site k's variable ℓ as

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

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

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 later provide a recommended version that builds on the following basic one. Formally, the basic version of SPS solves the following minimization problem.

$$\min_{(\mathbf{S},\ \mathbf{W})} \quad \frac{1}{N - N_S} \sum_{k=1}^{N} \underbrace{(1 - S_k)\Bigg(\frac{1}{L}\sum_{\ell=1}^{L} B_{k\ell}(\mathbf{W}, \mathbf{S})\Bigg)}_{\text{Imbalance for non-selected site } k} \tag{3}$$

with regular constraints that (i) the number of selected sites is N_S, and (ii) weights are positive and sum to one. In practice, we standardize variables 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 variables. Second, by multiplying this by $(1 - S_k)$, the objective function averages over the imbalance only for non-selected sites k. 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.

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, Diamond and Hainmueller, 2010; Xu, 2017).[5] When one wants to consider internal validity and choose a 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.

[5] While SPS is inspired by the synthetic control method, SPS does not presume data on a history of pre-treatment outcome variables.

Examples	**Formalization** (Add constraints to SPS)
Practical Constraints We cannot conduct studies in site k e.g., No online survey firm in some African countries	$S_k = 0$ 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 sites from different quantiles of GDP	$\sum_{k:S_k=1} \mathbf{1}\{\text{GDP}_k \leq 20 \text{ percentile}\} \geq 1$ $\sum_{k:S_k=1} \mathbf{1}\{\text{GDP}_k \in (40 \text{ and } 60 \text{ percentiles})\} \geq 1$ $\sum_{k:S_k=1} \mathbf{1}\{\text{GDP}_k \geq 80 \text{ percentile}\} \geq 1$

Table 3: **Examples of Constraints and Domain Knowledge that SPS can Incorporate.**

4.2.3 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 3 summarizes examples of domain knowledge and practical constraints users may add to SPS. The companion `R` package `spsR` allows users to incorporate these constraints using simple functions (see Appendix B).

First, researchers can easily incorporate practical, logistical and ethical constraints. 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 other cases, researchers might not be able to select certain countries because they do not have local collaborators or local knowledge. In these scenarios, users can add $S_k = 0$ for any infeasible sites k as a constraint, which guarantees that those infeasible sites are not selected. Similarly, if users want to always select a particular site, e.g., Uganda, as one of the study sites for its substantive importance, they can add $S_{\text{Uganda}} = 1$ as a constraint. When, for example, funding conditions require researchers to choose certain sites, they can also include such restrictions here.

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 two democracies and at least two autocracies. 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 countries from different quantiles of GDP. See formalization in Table 3.

Third, researchers can also incorporate various other domain knowledge into SPS. (a) Users can incorporate not only $\mathbf{X}_k$ themselves but also any flexible functions of site-level variables, e.g., interaction and higher order terms, to capture nonlinearity in the data. (b) Researchers can incorporate varying importance of site-level variables, e.g., based on predictive power as in the standard synthetic control method. (c) Users can ensure that selected sites are geographically diverse and distant enough from each other. Other examples include budget constraints, differential costs of each site, and different sample sizes in each site. See Appendix E.3 for how to formalize these different considerations within SPS.

Finally, 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.

$$\frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N} W_{jk} \underbrace{S_j(1-S_k)\frac{1}{L}\sum_{\ell=1}^{L}(X_{j\ell}-X_{k\ell})^2}_{\substack{\text{Distance between} \\ \text{Selected Site } j \text{ and Non-Selected Site } k}}, \tag{4}$$

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 increases efficiency of estimating the (subgroup) average-site ATEs.

$$\frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N} S_j(1-S_k)W_{jk}^2 \tag{5}$$

We provide formal discussions about these penalty terms in Section 5.2.

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 and subgroup average-site ATEs by combining causal estimates from selected sites. We also discuss how to empirically assess the potential influence of unobserved confounders using a procedure similar to cross-validation.

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.

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

$$\widehat{\theta}_{AS} := \frac{1}{N}\big(\sum_{j\in\mathcal{S}} \widehat{\theta}_j + \sum_{k\in\mathcal{R}} \widehat{\theta}_k^W\big) \tag{6}$$

where, for non-selected sites $k \in \mathcal{R}$, $\widehat{\theta}_k^W := \sum_{j\in\mathcal{S}} \widehat{W}_{jk}\widehat{\theta}_j$ and weights $\widehat{W}_{jk}$ are estimated in SPS (equation (3)). This simply averages over the site-specific ATE estimates from selected and non-selected sites. We emphasize that we primarily use $\widehat{\theta}_k^W$ as an intermediate step toward estimating the (subgroup) average-site ATE, which is the main quantity of interest defined in Section 4.1.

Similarly, the proposed SPS estimator for the subgroup average-site ATE is defined as,

$$\widehat{\theta}_{sub}^g := \frac{1}{N_g}\big(\sum_{j\in\mathcal{S},G_j=g} \widehat{\theta}_j + \sum_{k\in\mathcal{R},G_k=g} \widehat{\theta}_k^W\big) \tag{7}$$

where we average over the site-specific ATE estimates from selected and non-selected sites with variable G equal to g. For example, in Naumann, F. Stoetzer and Pietrantuono (2018), variable G could be fiscal exposure and $g \in \{\text{high, low}\}$. We propose the conservative variance estimator in Appendix E.4. We also discuss connections to and differences from conventional meta-analysis estimators in Appendix E.5.

The proposed SPS estimators are the optimal weighted average-based predictors that minimize the worst-case mean squared error (see Section 5.2). Based on this theoretical foundation, we only view the SPS estimator to be an optimal predictor given observed site-level variables, and importantly, we do not view 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 (subgroup) average-site ATE without (often infeasible) random sampling of sites, unless researchers make stringent modeling assumptions that we avoid in this paper. Rather, we focus on constructing estimators that can minimize the prediction error, while explicitly allowing for unobserved moderators.

Site-level Cross-Validation. 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.

For each iteration b, we randomly split selected sites into two equally sized sets, as-if-non-selected sites $\mathcal{S}_0^b$ and as-if-selected sites $\mathcal{S}_1^b$. Then, we test whether the difference between the average-site ATE estimates in the as-if-non-selected sites and the estimated average-site ATE based on the as-if-selected sites is statistically distinguishable from zero. Formally, the difference is defined as,

$$\widehat{\delta}_b := \underbrace{\frac{1}{N_{b,0}} \sum_{k \in \mathcal{S}_0^b} \widehat{\theta}_k}_{\substack{\text{ASATE in} \\ \text{As-If-Non-Selected Sites}}} - \underbrace{\frac{1}{N_{b,0}} \sum_{k \in \mathcal{S}_0^b} \widehat{\theta}_k^{W_b}}_{\substack{\text{ASATE estimated from} \\ \text{As-If-Selected Sites}}} \tag{8}$$

where $\widehat{\theta}_k^{W_b} = \sum_{j \in \mathcal{S}_1^b} \widehat{W}_{jk}^b \widehat{\theta}_j$ is a weighted average estimator for the ATE in the as-if-non-selected site k based on the as-if-selected sites $\mathcal{S}_1^b$. In each iteration b, we can obtain a p-value. As in typical cross-validation, we repeat the same procedure many times by randomly splitting selected sites and then combine p-values by the Holm–Bonferroni correction to account for multiple testing of dependent p-values.

When the difference is small and not statistically distinguishable from zero, there is no evidence of significant bias from unobserved site-level variables, while we can never confirm it as in usual statistical diagnostic tests. When the difference is large and statistically distinguishable from zero, 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 given multi-site 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 empirical application in Section 6.3).

In practice, when researchers expect relationships between observed and unobserved variables to be highly non-linear and they only have a few study sites, the statistical power of the test could be lower. In contrast, when observed and unobserved variables have stronger correlations or when researchers have more study sites, the statistical power is higher. Like any diagnostic test in causal inference, the site-level cross-validation is not a panacea, and thus, in practice, we recommend researchers augment the test with substantive discussion about potential unobserved variables (see Section 7.2). We also note that future studies can explore how to incorporate the equivalence testing approach (Hartman and Hidalgo, 2018) into the site-level cross-validation.

5.2 Statistical Properties

Formally, the proposed SPS estimator is the optimal weighted average-based predictor that minimizes the worst-case mean squared error.

We show that

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left\{\left(\theta_k - \widehat{\theta}_k(\mathbf{W})\right)^2\right\} \;\; &\lesssim \;\; \lambda_1 \times \frac{1}{N-N_S}\sum_{k=1}^{N}(1-S_k)\left(\frac{1}{L}\sum_{\ell=1}^{L}B_{k\ell}(\mathbf{W},\mathbf{S})\right) \\ &\quad + \lambda_2 \times \frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N}W_{jk}S_j(1-S_k)\frac{1}{L}\sum_{\ell=1}^{L}(X_{j\ell}-X_{k\ell})^2 \\ &\quad + \lambda_3 \times \frac{1}{N-N_S}\sum_{j=1}^{N}\sum_{k=1}^{N}S_j(1-S_k)W_{jk}^2. \qquad (9)\end{aligned}$$

where $\lesssim$ means that the inequality holds up to some constants unrelated to $(\mathbf{S},\mathbf{W})$, and $\widehat{\theta}_k(\mathbf{W}) := \sum_{j\in\mathcal{S}} W_{jk}\widehat{\theta}_j$ is a general weighted average estimator of the site-specific ATE. The first term on the right-hand side is exactly the same as the main objective function of the SPS algorithm (equation (3)), and the second and third terms are equivalent to the penalty terms in equations (4) and (5). $(\lambda_1, \lambda_2, \lambda_3)$ are some constant parameters that capture the relative importance of the three terms. Most importantly, because SPS directly minimizes the right-hand side of equation (9), the SPS estimator is a minimizer of the worst-case mean squared error.[6] We provide proof and additional discussions about the theoretical guarantees in Appendix E.1.

6 Empirical Applications

We now show that researchers can use SPS to systematically select diverse sites in a wide range of applications. We do this by reanalyzing three applications introduced in Section 2. Given space constraints, we offer two additional examples, including an application to an observational study, in Appendix A.

6.1 Multi-Country Survey Experiments on Immigration

We first illustrate SPS using Naumann, F. Stoetzer and Pietrantuono (2018), which uses a multi-country survey experiment—one of the most common types of multi-site experiments in recent years. In this study, the authors used survey experiments in 15 European countries to study whether and how much respondents prefer high-skilled immigrants to low-skilled immigrants (see Section 2.2).

We conduct empirical validation—pretending that we can only select a subset of sites that the original authors actually studied and validating whether we can recover the benchmark estimate of the (subgroup) average-site ATEs. In particular, we use SPS to select six sites out of 15 sites and then estimate the (subgroup) average-site ATE across all 15 sites. Because the original authors actually conducted experiments in all 15 sites, we can compare our SPS estimate based only on six sites to the actual experimental benchmark estimate. By doing so, we can simultaneously test the real-world performance of the method and illustrate the use of SPS.

[6] At the site selection stage, when we have not yet collected data, we cannot directly estimate the mean squared error itself. However, we can instead examine its upper bound, which incorporates the possibility of unmeasured moderators.

6.1.1 Site Selection

The first step is to specify the target population of sites against which we evaluate external validity. From this population of sites, SPS will purposively sample sites. The choice of the target population should be based on a given substantive theory of interest (see more discussions in Section 7). For the sake of a clear presentation, we use all 15 European countries in the original paper as the target population of sites.

The second step is to specify site-level variables to diversify. We include seven variables discussed in the original paper. The first four variables (GDP, size of migrant population, unemployment rates, and fiscal exposure) are country-level variables common in the immigration literature. Another variable is the baseline level of support for immigration by the general public (measured in previous waves of the European Social Survey). Finally, the last two variables (the mean age and the mean education) are country-level summary measures based on individual-level characteristics. These variables considered in the original paper are key site-level variables that are likely to explain the across-site heterogeneity of the ATEs.

The final step is to run SPS. As we recommend in Section 4, we include stratification to improve diversification: for each continuous variable, we make sure to select at least one site below the 20th percentile, at least one site between the 40th and 60th percentile, and at least one site above the 80th percentile. For a binary variable (i.e., fiscal exposure), we make sure to select at least one site with high exposure and at least one site with low exposure.

SPS selected Sweden, Denmark, Spain, Switzerland, Czechia, and the United Kingdom as six study sites. Figure 3 visualizes the results of SPS. To make visualization cleaner, we standardized each continuous 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 it is extremely difficult for humans to simultaneously diversify seven variables, SPS allows users to naturally diversify all chosen variables.

6.1.2 External Validity Analysis

Once experiments are conducted in each site, researchers can first report site-specific ATE estimates in the selected sites by focusing on internal validity (see Figure 4-(a)). Because SPS

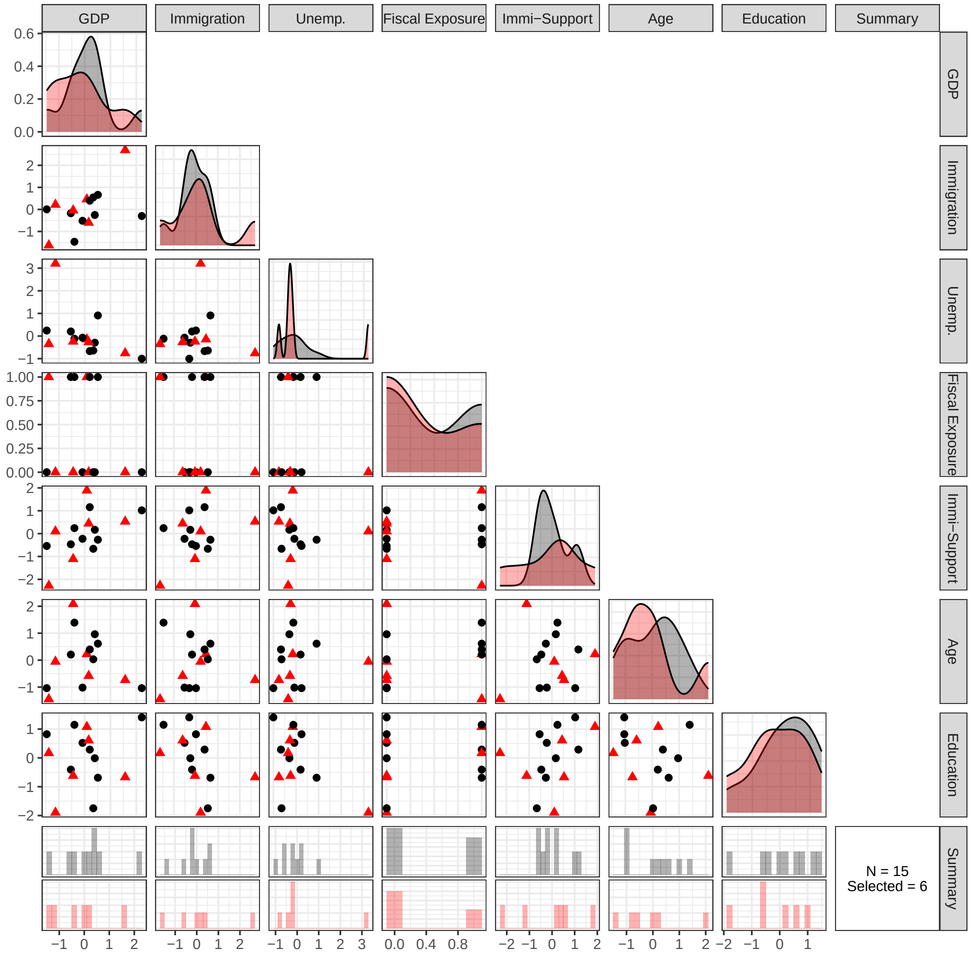


Figure 3: **SPS Site Selection for the Multi-Country Experiment on Immigration.** *Note*: In scatter plots, red triangles (black circles) represent values of site-level variables of selected (non-selected) sites. Plots in the last row and along the diagonal show the marginal distributions of each variable.

diversified the site selection, we see large heterogeneity even across the selected sites. The site-specific ATE estimates range from 13.5 percentage points to 41.3 percentage points.

For external validity analysis, we can first combine causal estimates from selected sites to estimate the average-site ATE. By using the SPS estimator (equation (6)), we estimate the average-site ATE to be 27.1 percentage points (95% CI = [15.1, 39.2]). Figure 4-(b) visualizes the results. In this empirical validation, we can explicitly compare our estimate to the actual experimental benchmark estimated from all 15 sites, which 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 lies within the 95% confidence interval. Second, as

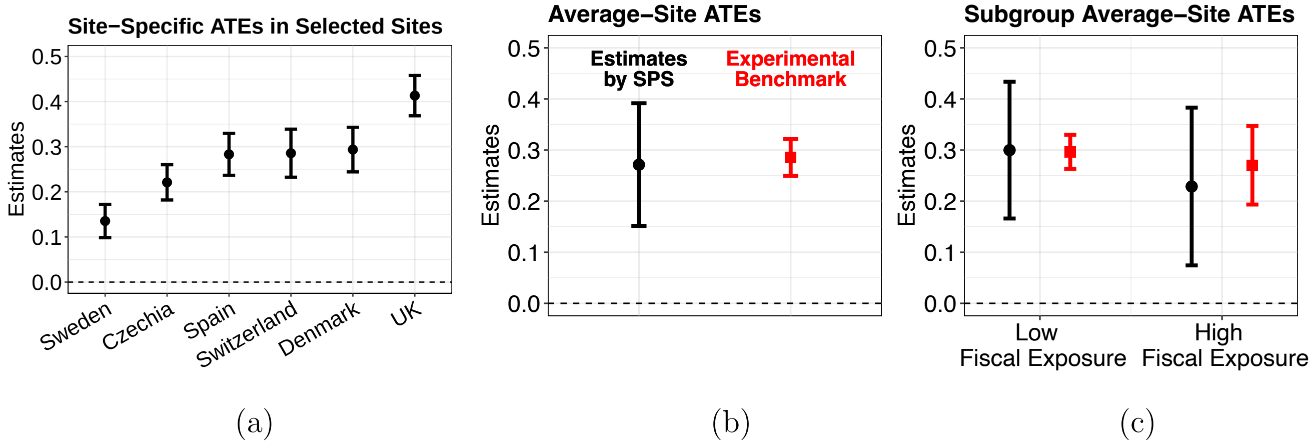


Figure 4: **Results of the Multi-Country Experiment on Immigration.**
Note: In Panels (b) and (c), the black circles and red squares represent estimates from SPS and the experimental benchmarks, respectively.

expected, the standard error of the SPS estimator based only on six 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.

Researchers can also explore the subgroup average-site ATEs. For example, in Naumann, F. Stoetzer and Pietrantuono (2018), researchers might be interested in testing whether the extent to which natives prefer high-skilled immigrants to low-skilled immigrants is stronger in countries with higher fiscal exposure to migration (i.e., the net burden of migration on public finances is higher) than in countries with lower fiscal exposure. To examine the implications of this hypothesis, we estimate the subgroup average-site ATEs for each subgroup (see Figure 4-(c)). In contrast to the theoretical prediction from the fiscal burden theory, we find that the extent to which natives prefer high-skilled immigrants to low-skilled immigrants is similar across countries with different levels of fiscal exposure to migration (findings consistent with Valentino et al. (2019)). We also find that our SPS estimates are close to the experimental benchmark.

Finally, it is recommended to investigate the potential influence of unobserved moderators using site-level cross-validation where we randomly choose three of the selected sites as if they were unobserved non-selected sites and predict the average ATE of those three non-selected sites based on the remaining three selected sites. We estimated the p-value to be 0.99, finding no evidence of significant bias from unobserved moderators, which is consistent with our comparisons against the experimental benchmark.

6.2 Metaketa Experiments on Community Policing

Blair et al. (2021) conducted a coordinated field experiment in six countries in the Global South to estimate the causal effect of community policing on crime and citizen-police relationships (see Section 2.3). As in many field experiments, this experiment by the Metaketa initiative was severely constrained by various practical constraints, e.g., funding conditions, whether local partners were willing to run experiments together with scientists, and whether researchers had knowledge of the local context and language. How can researchers use SPS to select diverse sites in settings where practical constraints are crucial, as in this study?

Following the original paper, we first define the target population of sites to be countries in the Global South. The original experiments only considered countries with moderate to large populations in Africa, Asia, and South America, so we also limit our target population to Global South countries with populations of size at least 1 million in the three regions. While we define the target population strictly based on publicly available information in the published paper, those with more private knowledge may define different target populations.

To choose site-level variables, we again closely follow the original paper (see page 4 of the original paper) and include eight moderators to diversify in SPS: regime type, freedom score, corruption score, criminal justice score, crime rate, the number of police personnel, Gini index, and GDP. These variables are selected by the original authors because they are theoretically expected to moderate the effectiveness of community policing. For example, in countries with higher levels of corruption, community policing that seeks to improve citizen-police relationships might not be effective due to already low levels of trust in police.

We explicitly incorporate practical constraints. To approximate realistic restrictions on the feasibility of experiments, we collected data on EGAP member countries and restricted SPS to select sites only from those countries. If researchers have other practical constraints, such as funding conditions or the availability of collaborators, they can incorporate such constraints as well. We also include several stratification conditions: (a) we choose two countries from each of the three regions, (b) we select at least two democracies and at least two autocracies, and (c) for

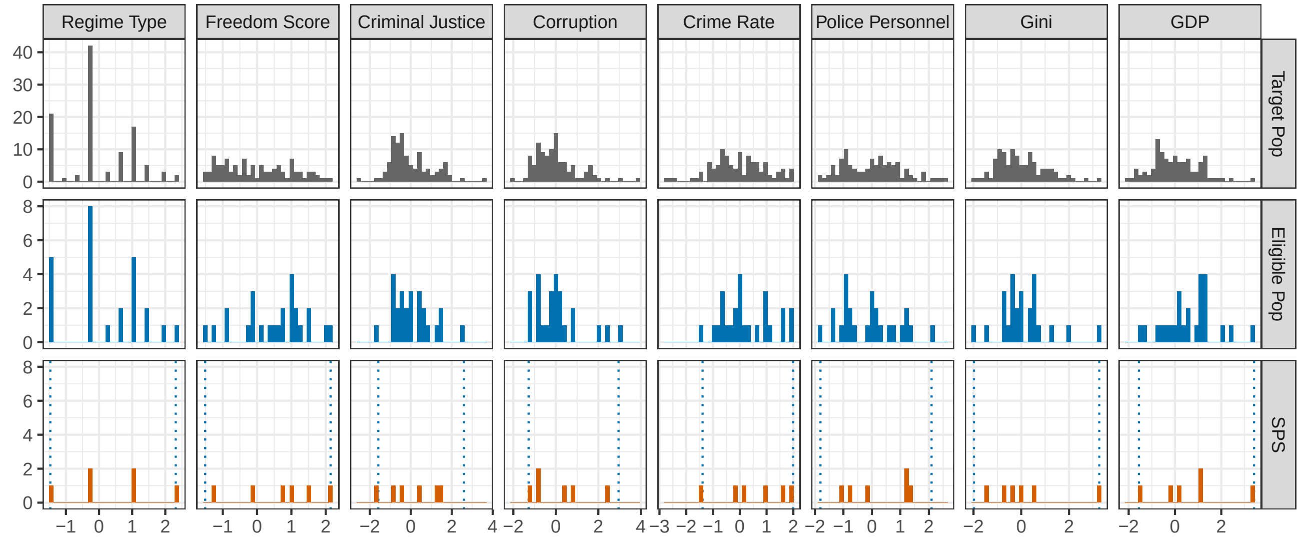


Figure 5: **SPS Site Selection for the Metaketa Experiment on Community Policing.** *Note*: For visual clarity, we standardized each variable.

the remaining variables, we select sites such that at least one selected site is above 1 standard deviation and at least one selected site is below -1 standard deviation. These stratification conditions further improve diversification by SPS.

SPS selected Bolivia, China, Liberia, Pakistan, South Africa, and Uruguay as six study sites where two sites (Liberia and Pakistan) overlap with the original site selection. Figure 5 compares the distributions of the eight moderators among the target population of countries in the Global South, the population of eligible sites that have EGAP members, and the SPS site selection. Several points are worth noting. First, SPS diversifies all eight moderators well, selecting countries with different regime types and with both low and high levels of criminal justice and corruption. Second, the SPS site selection is properly restricted to the EGAP member countries. For example, EGAP member countries do not include countries with extremely high or low levels of criminal justice and corruption, and as a result, SPS diversifies site selection within that constraint. This example illustrates how SPS selects diverse sites by effectively accounting for practical constraints often faced by researchers. In Appendix A, we compare the SPS site selection with the original site selection and show how SPS further improves diversity in site selection.[7]

[7] Note that we cannot report the (subgroup) average-site ATE estimates from the newly

6.3 Multi-County Field Experiments within the US

Gift and Gift (2015) conducted audit experiments in two US counties—one highly conservative and one highly liberal—to examine partisan bias in hiring (see Section 2.1). In this example, we illustrate how SPS can be used in combination with other site selection approaches, including the conventional purposive sampling and convenience sampling. In particular, we choose two additional sites complementary to the two sites selected by the original authors so that the four sites in total will jointly cover diverse contexts. This shows how researchers can first select some sites based on their choice of site selection methods—either theory-driven purposive sampling or non-theory-driven site selection required by practical constraints, such as funding conditions and availability of collaborators—and then use SPS to select remaining sites.

Based on the original paper, we begin by defining the target population of sites to be the US counties that are either highly liberal or highly conservative (counties whose vote share for a single party was greater than 60%, following the discussion in the original paper). To approximate the realism and feasibility of the large-scale audit experiment, we also limit the target population to counties with relatively large populations (at least 100K) and relatively high proportions of entry-level jobs for college graduates (above the median).

We follow the original paper to choose five site-level variables: Democrat vote share, unemployment rate, education level, population size, and rural population size. These variables are discussed by the original authors as key site-level variables that are theoretically expected to affect the extent of partisan bias. For example, we expect that partisan bias in hiring varies with unemployment rates because hiring companies in counties with higher unemployment rates are likely to have a bigger pool of applicants, which may lead companies to use partisanship for weeding out applicants who are otherwise similar in qualifications (see page 664 of the original paper).

We run SPS to select two additional sites to complement the original site selection. The goal

selected sites because the actual experiments were only conducted in the original six sites. When site-specific causal estimates are available from our selected sites, researchers can estimate the (subgroup) average-site ATE using our SPS estimator.

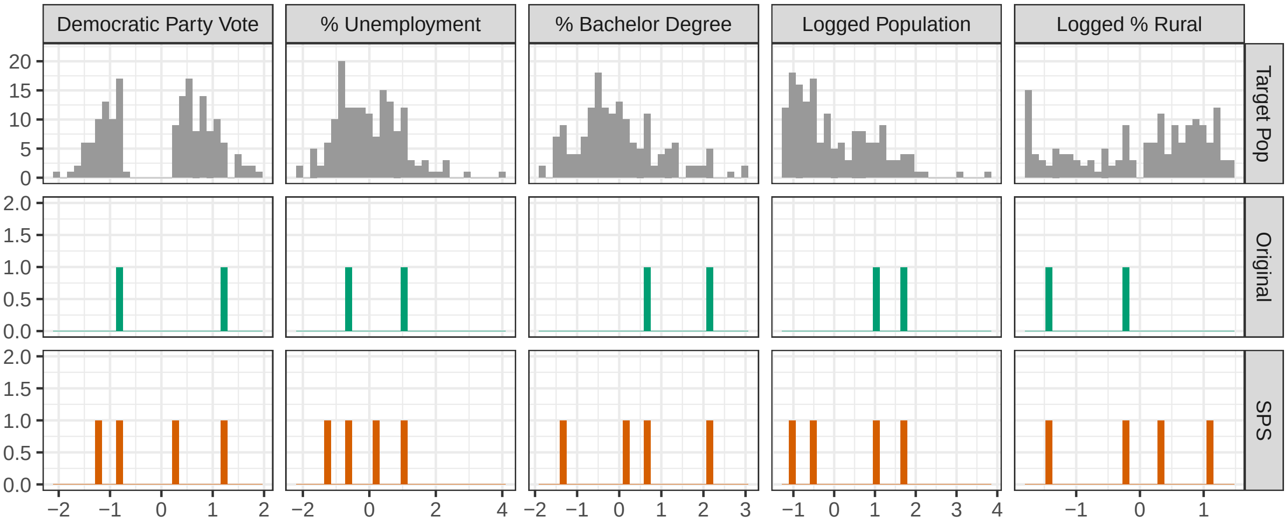


Figure 6: **SPS Site Selection for the Audit Experiment on Partisan Bias.**
Note: We standardized each variable. SPS (the last row) includes two additional sites as well as the original two sites.

is to select diverse study sites such that the two newly added sites can help researchers examine different types of sites that were not in the original selection. By doing so, the four sites in total will jointly cover a wide range of values in each site-level variable. As in other applications, to improve diversity, we include stratification to make sure that at least one selected site is above the 80th percentile and at least one selected site is below the 20th percentile of each variable.

SPS selected Blount County, TN, and Linn County, IA, to supplement the original site selection (Alameda County, CA, and Collin County, TX). Figure 6 compares the target population, the original site selection, and the SPS selection (four sites, including the original two sites). The original site selection diversified Democratic vote share and unemployment rate well, whereas study sites mostly focused on counties in populated urban areas with high levels of education. This original selection makes sense in establishing the first evidence of partisan bias in hiring. SPS can help accumulate knowledge for external validity by selecting a more diverse set of counties. In particular, SPS selected two additional sites that are complementary to the original site selection: two additional sites have lower proportions of people with Bachelor's degrees and smaller populations in more rural areas (see the third row in the figure). Combining them with the original site selection, we can cover a wide range of values in each of the five site-level variables.

7 Practical Guides

In this section, we provide practical recommendations and discuss precautions and limitations.

7.1 Defining the Target Population of Sites

In the first step of SPS, researchers need to define a population of sites of theoretical interest. This defines the target against which the external validity of a given substantive theory is evaluated. From this target population, SPS purposively selects study sites. Specifying the target population is similar to clarifying the studies' scope conditions, and thus, this choice should be guided by a given substantive research question.

Specification of the target population is essential because no causal finding is universally externally valid (Egami and Hartman, 2023); a study in a completely different context should, of course, return a different result. Therefore, explicit specification of the target population helps researchers guard against over-generalization. Specifying the target population of sites is important not only for SPS but also for any site selection approach that aims for better external validity (Findley, Kikuta and Denly, 2020). While this step has been typically implicit in the current practice of purposive sampling, our method makes this important step explicit.

In some settings, researchers might be interested in multiple sub-populations of sites rather than one population of sites. For example, Gift and Gift (2015) might be interested in testing whether the level of partisan bias differs across liberal and conservative counties in the US. In such cases, we recommend researchers explicitly define two sub-populations of sites based on the underlying theoretical mechanism—one for liberal and one for conservative counties.

7.2 Choosing Site-Level Variables

7.2.1 Avoid Including Irrelevant Variables

One of the benefits of SPS is that researchers can take into account a larger number of site-level variables. However, we recommend against a kitchen sink approach of including too many irrelevant variables because SPS might decrease the diversity in key site-level variables to improve the diversity in such irrelevant variables. In Appendix F, we use simulation studies to illustrate that including too many irrelevant variables indeed increases mean squared errors. Thus, we recommend researchers only include site-level variables that are substantively and theoretically expected to moderate treatment effects across contexts, as suggested in Bassan-Nygate et al. (2023).

7.2.2 How to Think about Unobserved Site-Level Variables

We clarify several points about how to reason about unobserved site-level variables.

SPS explicitly mitigates concerns about unobserved site-level variables compared to the current practice of purposive sampling. First, while researchers often only diversify one or two variables in current practice, users can include any number of site-level moderators in SPS based on their domain and theoretical knowledge. Second, systematically diversifying observed site-level variables can 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 larger (compared to site selection without SPS) only when diversifying observed site-level variables somehow *reduces* the diversity of unobserved site-level variables, which requires users to believe complicated nonlinear relationships between observed and unobserved variables.

While SPS can often mitigate concerns of unobserved site-level variables, it is recommended to empirically assess the influence of unobserved moderators using site-level cross-validation (see Section 5.1).

Finally, we emphasize that 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. SPS can reduce the mean squared error further if users can include more predictive moderators, but unobserved moderators do not invalidate the use of SPS.

7.3 Clarifications and Precautions

Combining SPS and Other Sampling Strategies. Researchers often have some domain knowledge or practical constraints that are 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. There are also practical (non-theory-driven) reasons, e.g., when multiple researchers from different countries collaborate, each investigator might need to include their own country of interest. One general approach to incorporate such additional information and constraints is to use SPS in combination with other site selection approaches, such as classical purposive sampling and convenience sampling. Using Gift and Gift (2015) as an example in Section 6.3, we illustrate how researchers can first select some sites based on their choice of site selection methods—either theory-driven purposive sampling or non-theory-driven site selection required by practical constraints, such as funding conditions and availability of collaborators—and then use SPS to select remaining sites.

Agnostic Use of SPS. Researchers might not be explicitly interested in estimating the (subgroup) average-site ATEs, 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 systematically diversify observed site-level covariates, while accommodating logistical and ethical constraints. 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.

Site-Hacking. It is important to advise *against* "site-hacking," i.e., re-running SPS until researchers select sites that they prefer, while justifying site selection as if it were selected without any additional constraint. For example, suppose researchers have local partners to run experiments only in three unrepresentative locations, but to justify their site selection, they decide to run SPS many times by post-hoc justifying different stratification conditions 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 and optimally diversify site selection within such constraints.

7.4 Limitations

SPS is not optimized for meta-regression. In multi-site causal studies, some researchers might be interested in running meta-regression with site-level variables. While this question is crucial, it is an even more difficult problem than estimating the average-site or the subgroup average-site ATE, which is already more challenging than internal validity problems. When the number of study sites is relatively small, as in political science, researchers have to estimate the effects of five site-level variables using only six 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, Pustejovsky and Ahmadi, 2019), whereas only 13% of multi-site studies in political science have more than 10 sites. Given this data constraint, most applications in political science have focused on the average-site and subgroup average-site ATEs, 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, but it is not an optimal approach. For those interested in meta-regression, we refer readers to Tipton, Pustejovsky and Ahmadi (2019).

8 Discussion

8.1 Connections to and Differences from Case Selection

While our focus is on 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; Glynn and Ichino, 2016; Nielsen, 2016). The qualitative case selection literature has developed a wide variety of sampling strategies, including typical, diverse, and extreme case selection, among others (e.g., Seawright and Gerring, 2008). In particular, the most common practice in multi-site quantitative studies is an instance of diverse case selection. Thus, SPS can also be seen as a hybrid of ideas from the 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, 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. Second, 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).

8.2 Other Dimensions of External Validity

Even though we focus on the external validity question about contexts, we emphasize the importance of other dimensions of external validity, such as treatments, outcomes, and populations (Findley, Kikuta and Denly, 2020; Egami and Hartman, 2023). In particular, recent papers emphasize issues such as sample representativeness (Mullinix et al., 2015; Coppock, Leeper and Mullinix, 2018), measurement harmonization (Slough and Tyson, 2022; Wilke and Samii, 2023), and the consequence of realistic and abstract treatments in survey experiments (Brutger et al., 2020).

Importantly, SPS can be useful for incorporating purposive variations in these other dimensions as well. For example, researchers can use SPS to select a set of treatments by diversifying implementation details, such as the program delivery model in field experiments (e.g., expert or volunteer canvassers) and the level of abstraction in survey experiments (e.g., use the real-world or hypothetical actors in vignettes).

9 Concluding Remarks

How should we select study sites for external validity? This has been a fundamental research design question for decades. For many quantitative social scientists, this question of site selection has recently become even more essential as an increasing number of scholars use multi-site causal studies to address external validity concerns about contexts (recall Figure 1).

This paper offers a new methodological foundation to design such increasingly popular, multi-site causal studies. SPS is a general method to select diverse sites for external validity in a wide range of applications. SPS can be used to select different types of sites, such as countries (for multi-country studies), cities, districts, states, and schools (for multi-site studies within a country). In general, SPS can be useful for selecting sites and cases when random sampling is infeasible due to practical constraints or ineffective due to the small number of sites researchers can select, which has been the case in most multi-site studies in political science. Researchers can implement all of the proposed methods with `R` package `spsR`.

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, and there are many valuable opportunities for scholars to develop and improve methodologies for multi-site studies. We hope that the proposed method in this paper can provide a useful foundation for future work.

References

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 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, Alberto and Jinglong Zhao. 2021. "Synthetic Controls for Experimental Design." *arXiv preprint arXiv:2108.02196* .

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

Arechar, Antonio A, Jennifer Allen, Adam J Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Michael N Stagnaro et al. 2023. "Understanding and Combatting Misinformation Across 16 Countries on Six Continents." *Nature Human Behaviour* 7(9):1502–1513.

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

Bassan-Nygate, Lotem, Jonathan Renshon, Jessica LP Weeks and Chagai M Weiss. 2023. "The Generalizability of IR Experiments Beyond the US.".

Blair, Graeme and Gwyneth McClendon. 2020. Experiments in Multiple Contexts. In *Handbook of Experimental Political Science*, ed. Donald P. Green and James Druckman. Cambridge University Press.

Blair, Graeme, Jeremy M Weinstein, Fotini Christia, Eric Arias et al. 2021. "Community Policing Does Not Build Citizen Trust in Police or Reduce Crime in the Global South." *Science* 374(6571):eabd3446.

Brutger, Ryan, Joshua D Kertzer, Jonathan Renshon, Dustin Tingley and Chagai M Weiss. 2020. "Abstraction and Detail in Experimental Design." *American Journal of Political Science* .

Chassang, Sylvain and Samuel Kapon. 2022. Designing Randomized Controlled Trials with External Validity in Mind. Technical report National Bureau of Economic Research.

Coppock, Alexander, Thomas J Leeper and Kevin J Mullinix. 2018. “Generalizability of Heterogeneous Treatment Effect Estimates Across Samples.” *Proceedings of the National Academy of Sciences* 115(49):12441–12446.

Devaux, Martin and Naoki Egami. 2022. “Quantifying Robustness to External Validity Bias.” *Available at SSRN 4213753* .

Doudchenko, Nick, Khashayar Khosravi, Jean Pouget-Abadie, Sebastien Lahaie, Miles Lubin, Vahab Mirrokni, Jann Spiess and Guido Imbens. 2021. “Synthetic Design: An Optimization Approach to Experimental Design with Synthetic Controls.” *Advances in Neural Information Processing Systems* 34:8691–8701.

Egami, Naoki and Erin Hartman. 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, Naoki and Erin Hartman. 2023. “Elements of External Validity: Framework, Design, and Analysis.” *American Political Science Review* 117(3):1070–1088.

Fearon, James D and David D Laitin. 2008. Integrating Qualitative and Quantitative Methods. In *The Oxford Handbook of Political Methodology*, ed. Henry E Brady, Janet Box-Steffensmeier and David Collier. Oxford University Press.

Findley, Michael G, Kyosuke Kikuta and Michael Denly. 2020. “External Validity.” *Annual Review of Political Science* .

Findley, Michael G, Michael Denly and Kyosuke Kikuta. 2023. *External Validity for Social Inquiry.*

Gechter, Michael, Keisuke Hirano, Jean Lee, Mahreen Mahmud, Orville Mondal, Jonathan Morduch, Saravana Ravindran and Abu S Shonchoy. 2023. “Site Selection for External Validity: Theory and an Application to Mobile Money in South Asia.”.

Gerber, Alan S and Donald P Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. WW Norton.

Gerring, John. 2006. *Case Study Research: Principles and Practices*. Cambridge university press.

Gift, Karen and Thomas Gift. 2015. "Does Politics Influence Hiring? Evidence from A Randomized Experiment." *Political Behavior* 37:653–675.

Glynn, Adam N and Nahomi Ichino. 2016. "Increasing Inferential Leverage in the Comparative Method: Placebo Tests in Small-n Research." *Sociological Methods & Research* 45(3):598–629.

Hartman, Erin and F. Daniel Hidalgo. 2018. "An Equivalence Approach to Balance and Placebo Tests." *American Journal of Political Science* 62(4):1000–1013.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12387*

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

McDermott, Rose. 2011. "Internal and External Validity." *Cambridge handbook of experimental political science* 27.

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics* 11(1):57–91.

Miratrix, Luke W, Michael J Weiss and Brit Henderson. 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.

Mullinix, Kevin J, Thomas J Leeper, James Druckman and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(2):109–138.

Munger, Kevin. 2019. "Knowledge Decays: Temporal Validity and SocialScience in a Changing World." *Working Paper* .

Naumann, Elias, Lukas F. Stoetzer and Giuseppe Pietrantuono. 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.

Nielsen, Richard A. 2016. "Case Selection via Matching." *Sociological Methods & Research* 45(3):569–597.

Olsen, Robert B, Larry L Orr, Stephen H Bell and Elizabeth A Stuart. 2013. "External Validity in Policy Evaluations that Choose Sites Purposively." *Journal of Policy Analysis and Management* 32(1):107–121.

Raudenbush, Stephen W and Xiaofeng Liu. 2000. "Statistical Power and Optimal Design for Multisite Randomized Trials." *Psychological methods* 5(2):199.

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

Shadish, William R, Thomas D Cook and Donald T Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Slough, Tara and Scott A Tyson. 2022. "External Validity and Meta-Analysis." *American Journal of Political Science* .

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

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

Tipton, Elizabeth, Larry Hedges, Michael Vaden-Kiernan, Geoffrey Borman, Kate Sullivan and Sarah Caverly. 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, Elizabeth and Laura R Peck. 2017. “A Design-based Approach to Improve External Validity in Welfare Policy Evaluations.” *Evaluation review* 41(4):326–356.

Tomz, Michael R and Jessica LP Weeks. 2013. “Public Opinion and the Democratic Peace.” *American Political Science Review* 107(4):849–865.

Valentino, Nicholas A, Stuart N Soroka, Shanto Iyengar, Toril Aalberg, Raymond Duch et al. 2019. “Economic and Cultural Drivers of Immigrant Support Worldwide.” *British Journal of Political Science* 49(4):1201–1226.

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

Wilke, Anna and Cyrus Samii. 2023. “To Harmonize or Not? Research Design for Cross-Context Learning.”.

Xu, Yiqing. 2017. “Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models.” *Political Analysis* 25(1):57–76.

Online Supplementary Appendix:

Designing Multi-Site Studies for External Validity:

Site Selection via Synthetic Purposive Sampling

Table of Contents

A Additional Empirical Applications

A.1 Multi-Country Survey Experiments: Lupu and Wallace (2019)

Lupu and Wallace (2019) examine conditions under which individuals are more likely to approve of human rights abuses by their governments. The original authors theorize that the approval rate varies by contexts involving the level of violence by both the government and the opposition as well as international legal constraints. The authors test their theory through multi-country survey experiments in India, Israel, and Argentina.

We use this example to demonstrate how researchers can optimally select study sites under a practical constraint common in multi-country survey experiments, i.e., accounting for the feasibility of conducting an online survey in each country.

A.1.1 Defining the Target Population

We closely follow the original paper to define the target population. First, the original authors selected all democratic countries where the levels of public approval of government are relatively more salient (Lupu and Wallace, 2019, p. 418). Second, the selected sites all have experienced significant opposition movements that are both violent and non-violent—an important condition in order for respondents to view the survey to be more realistic. Lastly, since one of the treatment factors was related to the government's compliance with international law, a country's relationship with international human rights institutions had to be salient in study sites. In the original paper, this concept was captured by whether countries joined the International Covenant on Civil and Political Rights (ICCPR) treaty. Based on these discussions in the original paper, we define the target population to be democratic countries that have experienced at least one opposition movement since 1970 and that have signed the ICCPR. The target population includes 58 countries in total: 16 in Americas, 18 in Europe, 15 in Africa, and 9 in Asia.

A.1.2 Choosing Site-Level Variables

We again follow the original authors' discussion on differences across the selected sites that potentially account for variation in attitudes toward law and violence. In particular, we choose the following seven site-level variables: polity score, civil liberty index, opposition group size, ethnic opposition group presence, population size, and the total number of treaties ratified out of the 18 human rights treaties under international law. The first two variables measure the extent to which democratic principles and civil liberty are respected in a given country, hence capturing varying levels of public approval of government and its saliency. The next two variables capture the predominance of opposition movement across all groups as well as ethnic groups in particular. These variables capture different relationships the government has with the opposition groups and, therefore, account for heterogeneity in the effect of the opposition movement on public approval. Population size is included to account for variations in demographic settings influencing the salience of government actions. The ratification of the international human rights treaties measures a country's compliance with international law, a key variable capturing baseline national attitude towards international institutions.

A.1.3 Stratification

Importantly, to capture practical concerns faced by researchers in an online survey setting, we incorporated practical constraints that researchers cannot run survey experiments in every country. In particular, we collected a list of countries where an online survey service is offered by four major survey platforms—Amazon Mechanical Turk, Lucid (now Cint), YouGov, and Dynata. We then restricted SPS to select sites only from these limited list of countries where online surveys are viable. This shows how researchers can use SPS to effectively incorporate practical constraints.

To improve diversification, we use stratification conditions: (a) For all but two variables, we select at least one site below the 25th percentile values, at least one site above the 75th, and at least one site between the 25th and 75th percentile values. (b) Due to the extreme asymmetry in the distributions of human rights treaties ratification as well as the size of ethnic opposition movement, we applied different stratification conditions to these two variables. Specifically, for the ratification measure, we used a stratification condition such that at least one selected site is below -1 and at least one selected site is above 1 standard deviation from the mean. Similarly, for the size of ethnic opposition movement, we use a stratification condition such that at least one selected site is below -0.5, at least one selected site is above 0.5, and at least one selected site between -0.5 and 0.5 standard deviation from the mean. As in this application, when the distribution of site-level variables is skewed, it is recommended to change stratification conditions flexibly for each variable in order to account for different patterns in each variable.

A.1.4 Site Selection

Given the practical constraints and stratification conditions specified above, we select three study sites, which is the same number of sites selected in the original paper. Specifically, SPS selected Bolivia, Kenya, and Lithuania as three study sites. Figure OA-1 shows the distribution of each site-level variable for the target population (gray), eligible countries where services are offered by the four survey providers (blue), sites selected by the original authors (green), and sites selected by SPS (orange). Several points are important. First, as specified by our practical constraint, the SPS site selection comes only from eligible countries. Second, we find that the sites selected by SPS cover a wider range of values in each of the site-level variables, especially for the opposition size, and the presence of ethnic opposition movement. For other variables, SPS is able to maintain similar or slightly better diversity than the original site selection.

The history of opposition movements against government regimes, and the government's responses, in Lithuania, Bolivia, and Kenya differ in several ways. In Lithuania, the opposition has largely been driven by political and social dissatisfaction, often channeled through peaceful protests and civil society organizations. The government has generally responded with a degree of tolerance for dissent, reflecting Lithuania's democratic institutions and respect for freedom of speech (Freedom House, 2022; U.S. Department of State, 2022c). In Bolivia, opposition movements have been marked by significant social and ethnic divisions. The country has experienced periods of political turmoil, including the ousting of Evo Morales in 2019, and the government has sometimes responded with force (U.S. Department of State, 2022a). In Kenya, brutality by government forces is a serious problem. Kenya has faced opposition movements related to both political and ethnic issues, often centered around contested

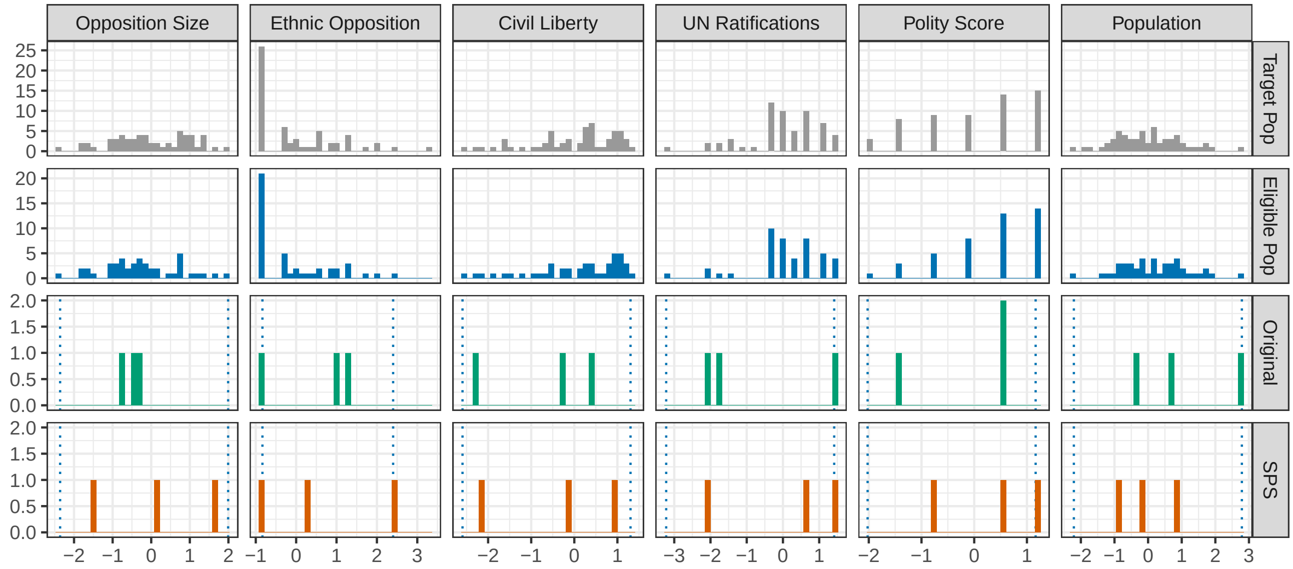


Figure OA-1: **Site Selection for Lupu and Wallace (2019).** *Note:* We compare the distribution of site-level variables among the target population (the first row), the eligible population (the second), the original selection (the third), and the SPS site selection (the fourth).

elections where the government often used excessive force to respond to the movement particularly when it is related to anti-government protests (U.S. Department of State, 2022b).

A.2 Multi-Context Observational Study: Bisbee *et al.* (2017)

In an influential observational study, Bisbee *et al.* (2017) examine the local average treatment effect (LATE) of fertility on labor supply using the same sex of the first two children as an instrumental variable (IV) based on the original design from Angrist and Evans (1998). The primary goal of this research is to assess the extent to which the quasi-experimental evidence on this effect found in a small number of countries in the previous literature—the US, Mexico, Argentina, and Taiwan—can be generalized to broader contexts. To do so, the original authors estimated the LATE of fertility on labor supply using exactly the same strategy in a wide range of countries over time, covering more than 40 countries over 50 years.

We will use this paper as an example to illustrate how researchers can also use SPS for observational studies. Similarly to how we used Naumann *et al.* (2018), we will conduct an empirical validation study based Bisbee *et al.* (2017). In particular, we pretend that we can only select a subset of sites that the original authors could actually study and then validate whether we can recover the benchmark estimate of the average-site ATE. We will use SPS to select 9 sites out of 45 sites and then estimate the average-site ATE of all 45 sites. Because the original authors used the instrumental variable method to estimate causal effects in all 45 sites, we can compare our SPS estimate based only on 9 sites to the actual quasi-experimental benchmark estimate. 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.

A.2.1 Defining the Target Population

To perform an empirical validation study, we start from the original authors' data set with 118 country-year pairs, containing 46 unique countries across the world—9 from Africa, 14 from Americas, 13 from Asia, and 10 from Europe. For reliable empirical validation, we remove pairs that have estimated LATEs with extremely large standard error values due to the problem of weak IV (e.g., Uganda in 2002 with a standard error over 2500; see Appendix A1 of Bisbee *et al.*, 2017). In particular, we remove estimates from pairs with standard errors greater than 2, which removes 7 pairs of estimates resulting in 45 unique countries. In this empirical validation, we use the 45 unique countries as our target population. The site-specific ATE is defined as the average effect within each country, averaging over time.

A.2.2 Choosing Site-Level Variables

We closely follow the country-level covariates included by the original authors: GDP per capita, female labor force participation, the sex ratio imbalance (male births per female births minus 0.5), and the total fertility rate. The first two variables capture the economic activity in each country that may influence the treatment effects. Furthermore, the original authors mention that the most significant negative IV estimates are among countries with higher levels of female labor force participation, indicating that the existing labor activity by women contributes to heterogeneity in the effect of fertility on labor supply. The third variable measures potential gender bias present in each country that is likely to capture the variation in treatment effects between different genders of mothers' first two children. The last variable is also a key variable capturing the baseline level of instrumental variable: The decision to have a third child based on preferences for sex heterogeneity is clearly less salient in countries where most families have more than three children.

In addition to the four variables discussed by the original authors, we also include two more variables that measure population size and education attainment. Population size functions in a similar manner as the total fertility rate in a sense that mothers in overly populated countries may behave differently than those in under-populated countries in terms of family planning. Education level is likely to capture the baseline level of employment rate. The original authors in fact include the education level for both mothers and spouses from the individual-level data set and find that complier mothers are more educated than the overall samples.

A.2.3 Stratification and Site Selection

In this application, we stratify our site selection such that at least two countries are selected from each of the four regions: Africa, Americas, Asia, and Europe. To showcase how researchers can combine the classical purposive sampling and SPS, we first select the United States (the most well-studied country in this topic) based on substantive importance and then select the remaining eight sites using SPS such that the nine sites jointly cover a wide range of values in each site-level variable we specify above. Finally, to further improve diversity, for each variable, we include stratification conditions such that we select at least two sites below the 20th percentile, at least two sites between the 40th and 60th percentiles, and at least two sites above the 80th percentile. SPS selected the following countries: Belarus, Chile, Costa Rica, India, Jordan, Rwanda, Spain, Uganda, and the United States. Figure OA-2 visualizes the result of SPS. We observe that SPS successfully diversified each variable, covering sites with smaller, closer to the mean, and larger values within each covariate.

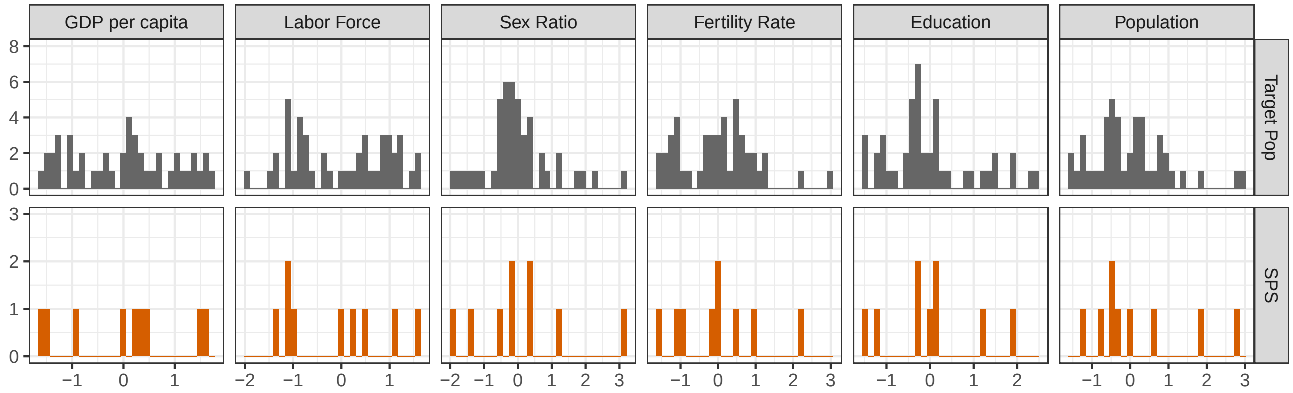


Figure OA-2: **Site Selection for Bisbee *et al.* (2017).**

A.2.4 Empirical Validation

Figure OA-3 shows the results of the site-specific LATEs in the selected sites (left panel), the average-site LATE (black coefficient plot in the right panel), as well as the benchmark estimate from all 45 countries (red coefficient plot in the right panel). Several points are worth discussing. First, as in the original paper, due to different strength of instrument across countries, we see that standard errors of site-specific LATEs differ widely across sites. Second, and most impotantly, we find that the point estimate from the proposed SPS based only on 9 countries closely resembles the benchmark estimate from all 45 countries and is within the 95% confidence interval. Finally, to assess the potential influence of unobserved confounders, we use the site-level cross-validation as recommended in the paper. The estimated p-value is over 0.99, and we did not find evidence for significant bias from unobserved confounders, which is consistent with our comparison against the benchmark estimate above.

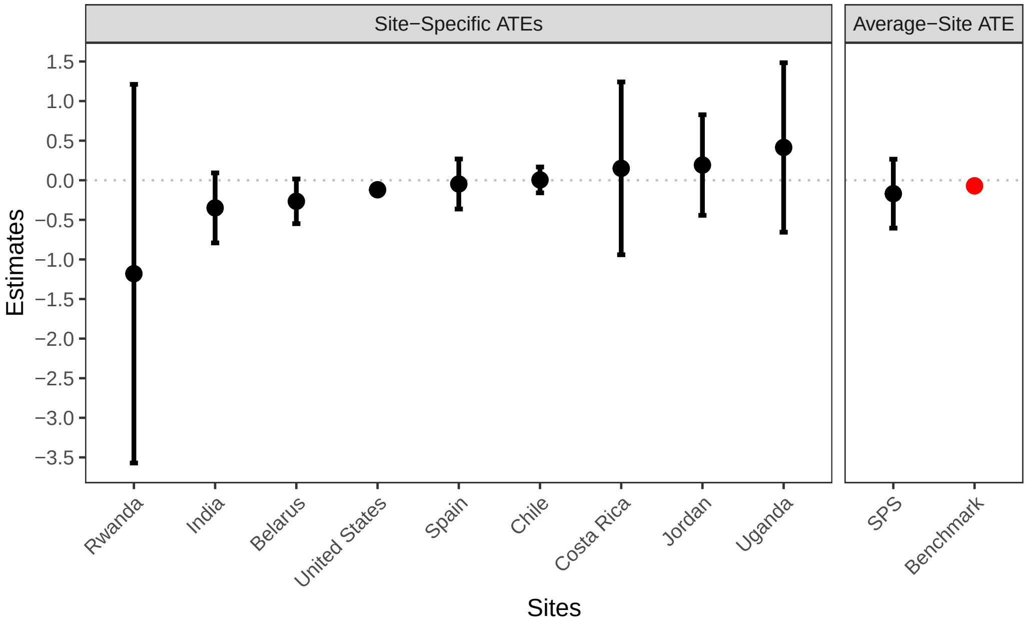


Figure OA-3: Site-Specific LATEs (left panel) and Average-Site LATE (right panel).

A.3 Metaketa Multi-Country Experiments: Blair *et al.* (2021)

In Section 6.2 of the main paper, we use Blair *et al.* (2021) and show how to use SPS to select diverse study sites within practical constraints. Here, we clarify how the SPS site selection differs from the original site selection.

Before we start, we clarify several key points. First, while we tried to approximate the target population and practical constraints based on public information in the published paper, the original authors had more private domain information and various logistical constraints that we could not incorporate. Therefore, our empirical application in Section 6.2 and in this section is not an evaluation or criticism of the original site selection. The only goal is to show how SPS can be effectively used in settings similar to the Metaketa study. Second, and most importantly, in future studies, researchers can incorporate any logistical constraints and domain knowledge (including constraints and theoretical considerations that were not explicitly documented in the original paper) explicitly in SPS, and thus, researchers designing the Metaketa experiments can also use SPS to further improve the transparency of the site selection process.

With this caveat, we now compare the SPS site selection and the original site selection in Figure OA-4. Several points are worth noting. First, the original site selection did diversify many key variables (e.g., crime rates, the number of police personnel) successfully, while they mostly focused on countries with lower levels of criminal justice and higher levels of corruption even within the Global South.[1] SPS diversified all eight variables successfully. Especially for the first four variables in Figure OA-4 (regime type, freedom score, criminal justice, and corruption), SPS significantly improved diversification compared to the original site selection. For the remaining four variables, we maintain the high level of diversity as in the original site selection.

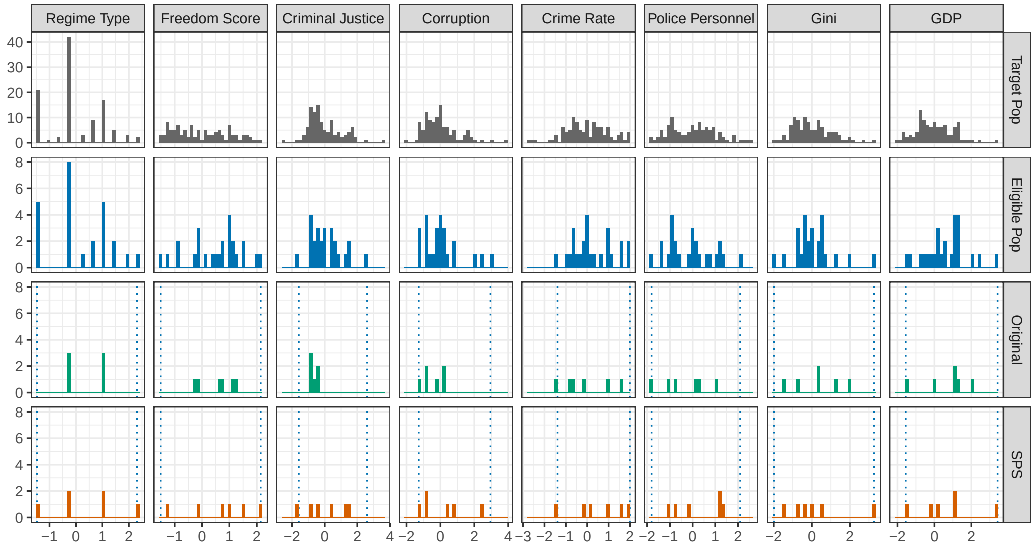


Figure OA-4: **Site Selection for Blair *et al.* (2021).** *Note:* We compare the distribution of site-level variables among the target population, the eligible population, and the original and SPS site selection.

[1] This site selection might make sense if the target population focuses only on countries in the Global South that have low levels of criminal justice and high levels of corruption.

A.4 Summaries of Site-Level Variables

In this section, we report the summary statistics of site-level variables for the two empirical applications in the paper. They show how SPS effectively diversify all the chosen variables within user-specified constraints.

Table OA-1: Multi-Country Survey Experiments: Naumann *et al.* (2018)

Site	GDP per capita (PPP in US$)	Migrant Stock (% of population)	Unemployment (% of labor force)	Fiscal Exposure	Support for Immigration	Respondent Age	Respondent Education
Selected Sites							
Sweden	45,297	14.76	8.05	1	4.01	49.85	2.84
Denmark	45,999	9.18	7.38	0	3.41	48.33	2.76
Spain	33,637	13.48	26.09	0	3.27	49.32	2.30
Switzerland	59,535	26.50	4.75	0	3.44	48.04	2.52
Czech Republic	31,187	3.79	6.95	1	2.28	46.70	2.68
UK	40,227	12.13	7.52	0	2.76	53.37	2.53
Min	31,187	3.79	4.75	0	2.28	46.70	2.30
Median	42,762	12.80	7.45	0	3.34	48.83	2.60
Mean	42,647	13.31	10.12	0.33	3.20	49.27	2.61
Max	59,535	26.50	26.09	1	4.01	53.37	2.84
Target Population							
Min	30,405	3.79	3.42	0	2.28	46.70	2.30
Median	45,297	12.13	7.52	0	3.27	49.49	2.68
Mean	44,498	12.34	8.83	0.40	3.22	49.43	2.65
Max	65,705	26.50	26.09	1	4.01	53.37	2.90
Non-Selected Sites							
Min	30,405	4.62	3.42	0	2.95	47.47	2.33
Median	46,393	11.43	8.19	0	3.13	49.82	2.70
Mean	45,733	11.70	7.96	0.44	3.24	49.53	2.67
Max	65,705	15.82	13.73	1	3.71	52.06	2.90

Note: Immigration Support is measured as the weighted average of sum of six scaled immigration attitude variables in the previous wave of European Social Survey (ESS). Age and Education represent average values among respondents in the author's original data.

Table OA-2: Metaketa Multi-Country Experiments: Blair *et al.* (2021)

Site	Regime	Freedom	Corruption	Criminal Justice	Crime Rate	Police Personnel	Gini	GDP
	Type	Score	Index	Index	(per 100K population)	(per 100K population)	Coefficient	(in 1M US$)
Selected Sites								
Bolivia	3	66	0.27	0.22	315.93	280.57	43.60	36,629
China	0	9	0.53	0.45	129.22	557.83	37.10	14,687,743
Liberia	6	60	0.32	0.31	87.09	176.51	35.30	3,039
Pakistan	3	37	0.31	0.35	22.78	203.12	29.60	300,425
Uruguay	9	96	0.73	0.56	573.94	590.26	40.20	53,666
South Africa	6	79	0.48	0.53	888.71	563.76	63	337,619
Min	0	9	0.27	0.22	22.78	176.51	29.60	3,039
Median	4.50	63	0.40	0.40	222.57	419.20	38.65	177,046
Mean	4.50	57.83	0.44	0.41	336.28	395.34	41.47	2,569,854
Max	9	96	0.73	0.56	888.71	590.26	63	14,687,743
Target Population								
Min	0	1	0.16	0.13	4.85	117.33	26	1,431
Median	3	37	0.40	0.36	100.01	304.98	38.60	37,605
Mean	3.50	40.31	0.42	0.40	182.80	342.70	39.98	282,715
Max	9	96	0.91	0.79	917.69	1, 141.61	63	14,687,743
Eligible Population								
Min	0	1	0.26	0.22	22.78	117.33	26	3,039
Median	3	59	0.41	0.38	102.99	289.29	39.50	121,347
Mean	3.96	53.60	0.43	0.41	227.29	322.20	40.53	891,954
Max	9	96	0.80	0.67	917.69	881.14	63	14,687,743
Non-Selected Sites								
Min	0	1	0.16	0.13	4.85	117.33	26	1,431
Median	3	36	0.40	0.36	97.71	304.98	38.60	35,432
Mean	3.43	39.25	0.42	0.39	173.50	339.51	39.89	144,101
Max	9	94	0.91	0.79	917.69	1, 141.61	59.10	2,671,595

Note: Regime Type is an ordinal variable measuring the political regime ranging from 0 (closed autocracy) to 9 (liberal democracy).

Table OA-3: Multi-County Field Experiments within the US: Gift and Gift (2015)

Site	Democratic Vote Share	Population	Rural Residency	Unemployment	Bachelor Degree
	(% of total votes)		(% of population)	(% of labor force)	(% of population)
Selected Sites					
Alameda County, CA	0.79	1,510,271	0.004	0.11	0.23
Linn County, IA	0.60	211,226	0.13	0.06	0.20
Blount County, TN	0.30	123,010	0.33	0.09	0.12
Collin County, TX	0.37	782,341	0.05	0.07	0.31
Min	0.30	123,010	0.004	0.06	0.12
Median	0.48	496,783	0.09	0.08	0.21
Mean	0.51	656,712	0.13	0.08	0.21
Max	0.79	1,510,271	0.33	0.11	0.31
Target Population					
Min	0.13	100,157	0	0.04	0.09
Median	0.61	246,310	0.12	0.09	0.18
Mean	0.54	540,435	0.15	0.09	0.19
Max	0.92	9,818,605	0.56	0.17	0.35
Non-Selected Sites					
Min	0.13	100,157	0	0.04	0.09
Median	0.61	246,310	0.12	0.09	0.18
Mean	0.54	537,491	0.15	0.09	0.19
Max	0.92	9,818,605	0.56	0.17	0.35

Note: Democratic Vote Share represents a share of votes received by a Democratic candidate in the 2008 presidential election.

Table OA-4: Multi-Country Survey Experiments: Lupu and Wallace (2019)

Site	Polity Score	Civil Liberty Index	Opposition Group Index	Opposition Ethnic Group Index	UN Ratification	Population (in 1,000s)
Selected Sites						
Bolivia	7	0.85	-0.23	0	18	11,090
Kenya	9	0.68	1.81	0.78	8	46,851
Lithuania	10	0.94	-2.31	0.29	16	2,904
Min	7	0.68	-2.31	0	8	2,904
Median	9	0.85	-0.23	0.29	16	11,090
Mean	8.67	0.82	-0.24	0.35	14	20,282
Max	10	0.94	1.81	0.78	18	46,851
Target Population						
Min	5	0.64	-3.49	0	5	330
Median	8.50	0.88	-0.47	0.14	14	11,323
Mean	8.17	0.86	-0.37	0.20	14.03	58,149
Max	10	0.96	2.27	1	18	1,322,866
Eligible Population						
Min	5	0.64	-3.49	0	5	330
Median	9	0.89	-0.79	0.14	14	17,033
Mean	8.61	0.86	-0.70	0.18	14.25	74,247
Max	10	0.96	2.27	0.78	18	1,322,866
Non-Selected Sites						
Min	5	0.64	-3.49	0	5	330
Median	8	0.88	-0.47	0.14	14	11,557
Mean	8.15	0.86	-0.37	0.19	14.04	60,215
Max	10	0.96	2.27	1	18	1,322,866

Note: Opposition Group Index measures the size opposition actors to the current political regime. Opposition Ethnic Group Index measures the size of active racial/ethnic group that mobilize against the political regime. UN Ratification refers to the total number of treaties ratified by a country out of the 18 human rights treaties under international law.

Table OA-5: Multi-Context Observational Study: Bisbee *et al.* (2017)

Site	GDP per capita	Female Labor Force	Sex Ratio	Fertility Rate	Female Educational	Population
	(in US$)	(% female adult population)	Imbalance	(per woman)	Attainment	(in 1,000s)
Selected Sites						
Belarus	5,678	0.82	0.52	1.76	65.54	10,026
Chile	5,901	0.31	0.51	2.61	33.37	12,584
Costa Rica	7,625	0.34	0.51	2.74	31.92	3,069
India	1,398	0.31	0.54	2.70	37.21	875,081
Jordan	3,946	0.27	0.52	4.12	39.79	5,532
Rwanda	731	0.92	0.50	3.39	12.66	8,372
Spain	23,524	0.56	0.52	2.01	39.75	39,897
Uganda	582	0.68	0.50	3.09	6.22	18,171
United States	27,324	0.62	0.51	2.16	80.28	236,553
Min	582	0.27	0.50	1.76	6.22	3,069
Median	5,678	0.56	0.51	2.70	37.21	12,584
Mean	8,523	0.54	0.51	2.73	38.53	134,365
Max	27,324	0.92	0.54	4.12	80.28	875,081
Target Population						
Min	582	0.13	0.50	1.72	4.98	1,897
Median	4,778	0.58	0.51	2.74	33.25	15,404
Mean	7,804	0.57	0.51	2.74	38.05	69,779
Max	32,269	0.92	0.54	4.63	89.97	1,070,038
Non-Selected Sites						
Min	726	0.13	0.50	1.72	4.98	1,897
Median	4,680	0.61	0.51	2.87	32.21	16,361
Mean	7,624	0.57	0.51	2.74	37.93	53,633
Max	32,269	0.92	0.53	4.63	89.97	1,070,038

Note: Sex Ratio Imbalance is measured as the number of male children divided by the number of female children minus 0.5. Female Educational Attainment is measured as the share of female population ages 25+ with at least lower secondary education.

B Introduction to R Package

In this section we provide a brief introduction to our companion `R` package `spsR`. We demonstrate the use of functions in the package using the empirical application of Naumann *et al.* (2018) described in the main manuscript.

We begin with the data set of seven site-level variables collected for the target population of 15 European countries included in the original study.

```
head(X)
```

```
##                      GDP Immigration      Unemp. Fiscal Exposure Immi-Support
## Netherlands  0.39258417  -0.2515705 -0.29438111               0    0.1686590
## Sweden       0.08340981   0.4598595 -0.14409826               1    1.8877250
## Norway       2.21404326  -0.2987159 -1.00312205               0    1.0183436
## France      -0.53687041  -0.1733010  0.20099546               1   -0.4658665
## Germany      0.19781674   0.3966613 -0.66730499               1    1.1578541
## Belgium     -0.08361722  -0.5149497 -0.07359521               0   -0.2262751
##                     Age   Education
## Netherlands  0.9612745 -0.01125098
## Sweden       0.2240878  1.07703001
## Norway      -1.0382000  1.40463468
## France       0.2073008 -0.40711759
## Germany      0.3962881  0.28840065
## Belgium     -1.0186468  0.52159260
```

As recommended in Section 7, researchers can include stratification conditions to improve diversity. To do so, users can rely on the function `stratify_sps()`. As described in the main manuscript, we apply a stratification for all but fiscal exposure variables such that the SPS selects at least one site below the 20th percentile, at least one site between the 40th and 60th percentile, and at least one site above the 80th percentile. Users can use loop to apply the same stratification condition for every variable. For the binary measure of fiscal exposure, we apply stratification such that the SPS selects at least one country with a value of 1 and one country with a value of 0.

```
col_div <- setdiff(colnames(X_use), "Fiscal Exposure")

st_list_btw <- st_list_large <- st_list_small <- list()
for (i in col_div){
  st_list_small[[i]] <- stratify_sps(X = X_use, num_site = list('at least', 1),
                                     condition = list(i, 'smaller than or equal to',
                                                      quantile(X_use[,i], probs = 0.2)))
  st_list_large[[i]] <- stratify_sps(X = X_use, num_site = list('at least', 1),
                                     condition = list(i, 'larger than or equal to',
                                                      quantile(X_use[,i], probs = 0.8)))
  st_list_btw[[i]] <- stratify_sps(X = X_use, num_site = list('at least', 1),
                                   condition = list(i, 'between',
                                                    quantile(X_use[,i], probs = c(0.4, 0.6))))
}
```

```
st_list_small$"Fiscal Exposure" <-
  stratify_sps(X = X_use, num_site = list('at least', 1),
               condition = list("Fiscal Exposure", 'smaller than or equal to', 0.5))
```

```
st_list_large$"Fiscal Exposure" <-
  stratify_sps(X = X_use, num_site = list('at least', 1),
               condition = list("Fiscal Exposure", 'larger than or equal to', 0.5))
```

Given the data and user-specified stratification conditions, users can run the function `sps()` to perform SPS. Note that, if users do not want to include any stratification condition, they can simply run the function `sps()` without specifying the argument `stratify` below.

```
out <- sps(X = X_use, N_s = 6, stratify = c(st_list_large, st_list_small, st_list_btw))
```

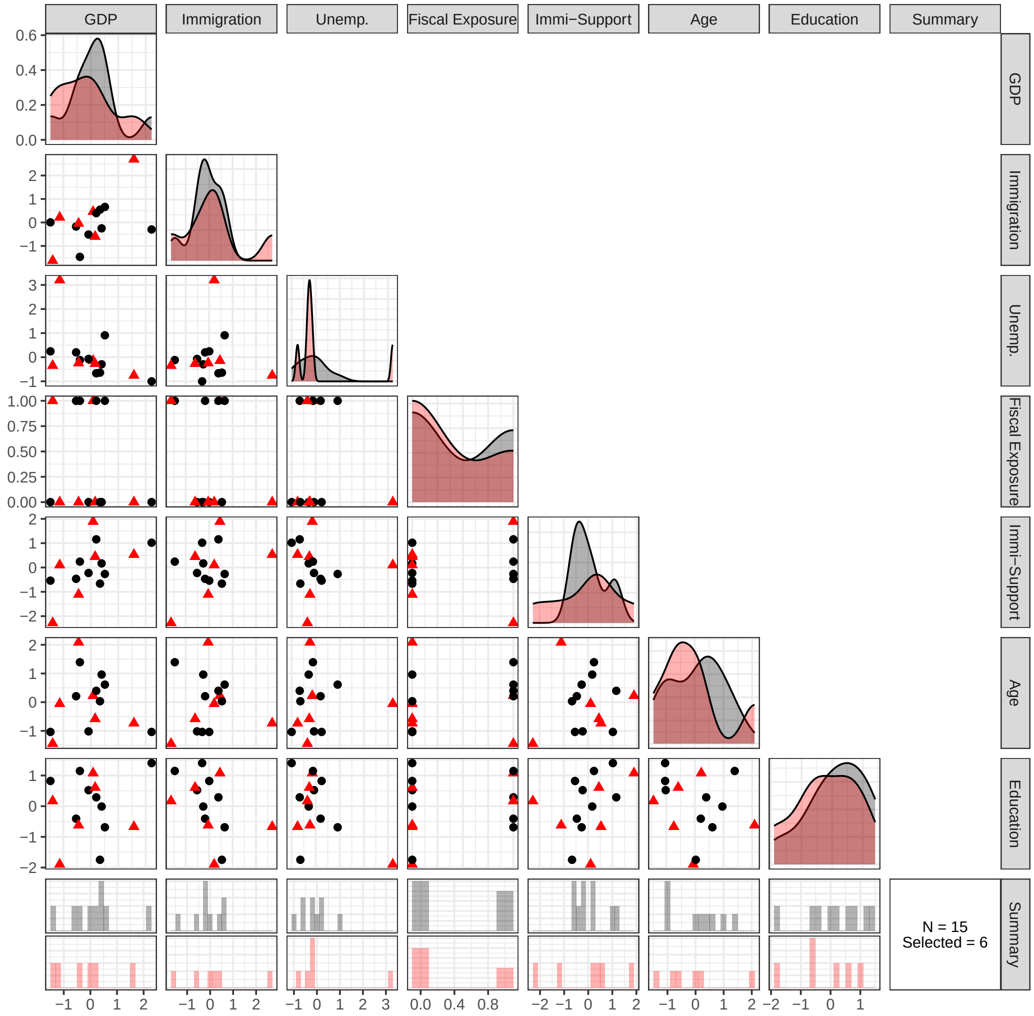
Selected sites are stored in an object called `selected_sites`:

```
out$selected_sites
```

```
## [1] "Sweden"         "Denmark"        "Spain"          "Switzerland"
## [5] "Czech Republic" "UK"
```

Users can also visualize the SPS selection along with the distribution of site-level variables using the function `sps_plot()`:

```
sps_plot(out)
```



Next, once the studies are conducted and the treatment effects are estimated in the selected sites, users can combine the results to estimate the average-site ATE:

```
# Estimated ATEs in Selected Sites
head(estimates_selected_sites)
```

```
##                       coef         se      p-value
## Sweden         0.1353307 0.01890234 1.179995e-12
## Denmark        0.2936739 0.02520817 5.692233e-30
## Spain          0.2831690 0.02368973 1.365971e-31
## Switzerland    0.2857338 0.02715871 9.747932e-25
## Czech Republic 0.2210802 0.01993582 8.498902e-28
## UK             0.4132192 0.02276622 4.294257e-67
```

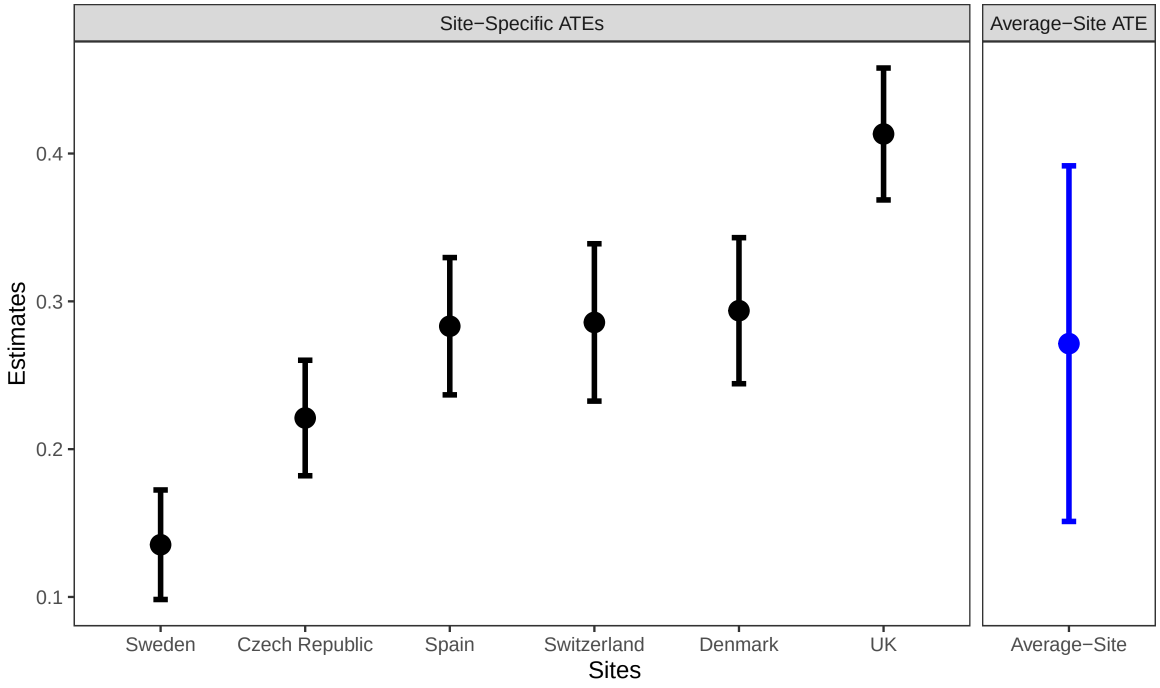
Using the estimated ATEs in selected sites, users can run the function `sps_estimator()` to estimate the average-site ATE:

```
# Estimating Average-Site ATE
sps_est <- sps_estimator(out = out, estimates_selected = estimates_selected_sites)
summary(sps_est)
```

```
##
##   Estimate Std. Error  CI Lower  CI Upper      p value
##  0.2713865 0.06137222 0.1510992 0.3916739 4.890107e-06 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Users can run `plot` to visually compare the average-site ATE against the site-specific ATEs:

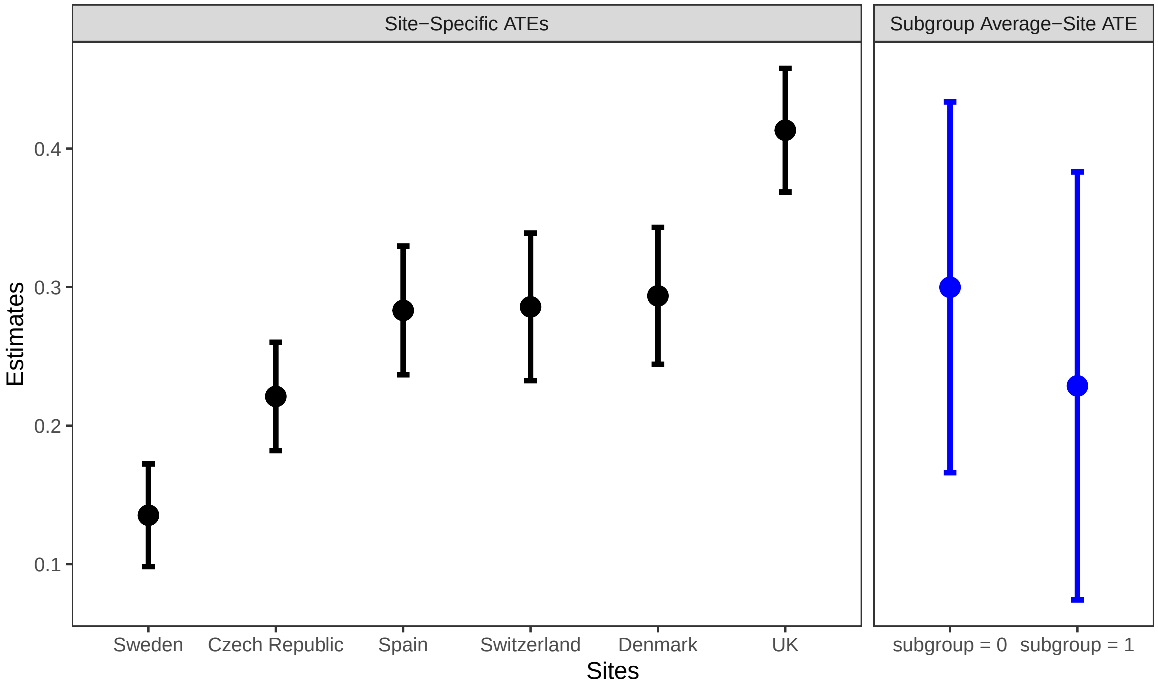
```
plot(sps_est)
```



Users can also estimate the subgroup average-site ATEs for each subgroup of interest using the same function:

```
sps_est_sub <- sps_estimator(out = out, estimates_selected = estimates_selected_sites,
                             subgroup = X[, "Fiscal Exposure"])
summary(sps_est_sub)
```

```
##
## Average-Site ATE:
##   Estimate Std. Error  CI Lower  CI Upper      p value
##  0.2713865 0.06137222 0.1510992 0.3916739 4.890107e-06 ***
##
## Subgroup Average-Site ATE:
##               Estimate Std. Error   CI Lower  CI Upper      p value
## subgroup = 0 0.2998686 0.06827125 0.16605944 0.4336778 5.607582e-06 ***
## subgroup = 1 0.2286870 0.07880572 0.07423067 0.3831434 1.854480e-03  **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Lastly, users can perform the site-level cross-validation as described in the main manuscript using the function `sps_cv()`:

```
sps_cv_out <- sps_cv(out = out, estimates_selected = estimates_selected_sites)
sps_cv_out$p_value
```

```
## [1] 0.997689
```

C Practical Guides on Collecting Site-Level Data

SPS requires site-level data to explicitly diversify site-level variables. In this section, we provide information on publicly available data that researchers may collect to perform SPS in a wide range of settings.

First, the majority of multi-site studies published to this date are conducted at the country-level. Specifically, across all 133 multi-site studies we analyzed (see Appendix D), 69% are multi-country studies. We emphasize that there exists ample amount of country-level data sets that are publicly available: a number of international organizations collect comprehensive measures of country's demographic, social, economic, and political indicators (e.g., UN Data, World Bank, WHO). A number of non-profit organizations and research institutes also collect a specialized set of measures at the country-level (e.g., V-Dem, Freedom House, World Justice Project) and an increasing number of academic groups are making collective efforts towards compiling a unique set of cross-country data for topics including, but not limited to, anti-government movement (NAVCO Data Project, Carnegie Global Protest Tracker) and UN General Assembly voting data (Voeten *et al.*, 2009). Furthermore, there exists public opinion survey data at both regional and global level (e.g., Gallup World Poll, European Social Survey, Eurobarometer, Afrobarometer, Americas Barometer, Asian Barometer). These measures are easily retrievable online via a direct download and many of the providers also offer API. Lastly, to facilitate the ease of cross-country data collection, our open-source `R` package `spsRdata` provides a country-level data set for 179 countries between 2010 and 2022, which includes a wide collection of relevant political, economic, and demographic indicators from V-Dem and the World Bank.

The next most prominent location in which multi-site studies are conducted is the United States. In particular, for the remaining 31% of the multi-site studies we analyzed (that are not multi-country), we find that 68% of them are conducted in the United States at the state (68% of multi-site studies in the United States), city (18%) or county (11%) level. Public data on United States abound, including government sources (Census of Governments, U.S. Bureau of Labor Statistics, FBI Uniform Crime Reporting), public opinion surveys (e.g., ANES, CCES, GSS), and an array of specific indicators developed by research institutes and advocacy groups including, but not limited to, election data and various social indicators (e.g., MIT Election Lab, ICPSR, ACLED). Furthermore, many of the replication materials from prior academic research conducted in the U.S. provides unique measures not commonly found elsewhere (e.g., Harvard Dataverse, de Benedictis-Kessner *et al.*, 2022).

For multi-site studies outside of the U.S., the country of interest often has statistics provided by the government as well as country-specific survey data available online. For example, Blair *et al.* (2021) collect data on police capacity from the selected countries' Census data as well as government annual reports. Lyall *et al.* (2015) use village-level data conducted by Opinion Research Center of Afghanistan (ORCA), an Afghan-owned firm that recruits its enumerators from sampled and neighboring villages.

D Literature Review of Multi-Site Causal Studies

D.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.[2]

D.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 1,337 articles. We then classified whether the experiment discussed in each article is a multi-site study using a two-step approach combining GPT-labeling and experts-manual-verification. First, we used GPT to label each article as a multi-site experimental study based on the article abstract. To increase accuracy, we used few-shot learning by inserting six abstracts and corresponding answers prior to providing an abstract of interest. GPT classified a total of 147 articles as a multi-site experiment. In the second step, we then manually coded the 147 articles that were labeled as a multi-site by GPT as well as a random selection of 146 articles that were labeled as a non-multi-site by GPT. In this verification step, we coded a total of 111 articles as multi-site out of the 293 articles reviewed: 97 out of the 147 articles labeled as a multi-site experiment by GPT were verified as such, and 14 out of the 146 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 likely the lower bound of the true number of multi-site studies that exist during the time frame we examine.

Tables OA-6 through OA-8 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 both field and survey multi-site experiments).

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

Table OA-6: 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 et al. (2012; AJPS)	Inequity Aversion and the International Distribution of Trade Protection
5	Lyall et al. (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
7	Aaroe and Petersen (2014; JOP)	Crowding Out Culture: Scandinavians and Americans Agree on Social Welfare in the Face of Deservingness Cues
8	Ocantos et al. (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	Mccauley (2014; APSR)	The Political Mobilization of Ethnic and Religious Identities in Africa
11	Bloom et al. (2015; APSR)	Religious Social Identity, Religious Belief, and Anti-Immigration Sentiment
12	Lyall et al. (2015; JOP)	Coethnic Bias and Wartime Informing
13	Carnes and Lupu (2016; APSR)	Do Voters Dislike Working-Class Candidates?
14	Lu and Scheve (2016; CPS)	Self-Centered Inequity Aversion and the Mass Politics of Taxation
15	Zink and Dawes (2016; PB)	The Dead Hand of the Past? Toward an Understanding of Constitutional Veneration
16	Bechtel and Scheve (2017; JEPS)	Who Cooperates? Reciprocity and the Causal Effect of Expected Cooperation in Representative Samples
17	Findley et al. (2017; JOP)	External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation
18	Gschwend et al. (2017; JOP)	Weighting Parties and Coalitions: How Coalition Signals Influence Voting Behavior
19	Laustsen (2017; PB)	Choosing the Right Candidate
20	Soroka et al. (2017; JEPS)	Ethnoreligious Identity, Immigration, and Redistribution
21	Wright et al. (2017; CPS)	Multiculturalism and Muslim Accommodation: Policy and Predisposition Across Three Political Contexts
22	Auerbach and Thachil (2018; APSR)	How Clients Select Brokers: Competition and Choice in India's Slums
23	Carlin and Love (2018; BJPS)	Political Competition, Partisanship and Interpersonal Trust in Electoral Democracies
24	Sheffer et al. (2018; APSR)	Nonrepresentative Representatives: An Experimental Study of the Decision Making of Elected Politicians
25	Lee (2019; CPS)	The Revival of Charisma: Experimental Evidence From Argentina and Venezuela
26	Bisgaard (2019; AJPS)	How Getting the Facts Right Can Fuel Partisan-Motivated Reasoning
27	Frye et al. (2019; WP)	Vote Brokers, Clientelist Appeals, and Voter Turnout: Evidence from Russia and Venezuela
28	Lupu and Wallace (2019; AJPS)	Violence, Nonviolence, and the Effects of International Human Rights Law
29	Kenan and Zohlnhoefer (2019; PB)	Policy and Blame Attribution: Citizens' Preferences, Policy Reputations, and Policy Surprises
30	Valentino et al. (2019; BJPS)	Economic and Cultural Drivers of Immigrant Support Worldwide
31	Chen and MacDonald (2020; JEPS)	Bread and Circuses: Sports and Public Opinion in China
32	Chilton et al. (2020; BJPS)	Reciprocity and Public Opposition to Foreign Direct Investment
33	Goerres et al. (2020; BJPS)	What Makes People Worry about the Welfare State? A Three-Country Experiment
34	Jensen and Rosas (2020; JEPS)	Open for Politics? Globalization, Economic Growth, and Responsibility Attribution
35	Mutz and Lee (2020; APSR)	How Much is One American Worth? How Competition Affects Trade Preferences
36	Tomz and Weeks (2020; JOP)	Human Rights and Public Support for War
37	Tomz et al. (2020; IO)	Public Opinion and Decisions About Military Force in Democracies
38	Avdagic and Savage (2021; BJPS)	Negativity Bias
39	Pereira (2021; PB)	Do Female Politicians Face Stronger Backlash for Corruption Allegations?
40	Blais and Vallve (2021; PB)	Conformity and Individuals' Response to Information About Aggregate Turnout
41	Bush and Zetterberg (2021; AJPS)	Gender Quotas and International Reputation
42	Dellmuth and Tallberg (2021; BJPS)	Elite Communication and the Popular Legitimacy of International Organizations
43	Doces and Wolaver (2021; PB)	Are WeAllPredictably Irrational? An Experimental Analysis
44	Edwards and Arnon (2021; BJPS)	Violence on Many Sides: Framing Effects on Protest and Support for Repression
45	Freire et al. (2021; JEPS)	Institutional Design and Elite Support for Climate Policies: Evidence from Latin American Countries
46	Goodman (2021; CPS)	Immigration Threat, Partisanship, and Democratic Citizenship: Evidence from the US, UK, and Germany
47	Hubscher et al. (2021; BJPS)	Voter Responses to Fiscal Austerity
48	Incerti et al. (2021; BJPS)	Hawkish Partisans: How Political Parties Shape Nationalist Conflicts in China and Japan
49	Kitagawa and Chu (2021; WP)	The Impact of Political Apologies on Public Opinion
50	Klasnja et al. (2021; JEPS)	When Do Voters Sanction Corrupt Politicians?
51	Magni and Reynolds (2021; JOP)	Voter Preferences and the Political Underrepresentation of Minority Groups
52	Robison et al. (2021; CPS)	Does Class-Based Campaigning Work? How Working Class Appeals Attract and Polarize Voters
53	Wood et al. (2021; JEPS)	The Effect of Geostrategic Competition on Public Attitudes to Aid
54	Yu et al. (2021; PB)	The (Null) Effects of Happiness on Affective Polarization, Conspiracy Endorsement, and Deep Fake Recognition
55	Aarslew (2022; BJPS)	Why Don't Partisans Sanction Electoral Malpractice?

Table OA-6: Multi-site Survey Experiments

N	Author (Year; Journal)	Title
56	Arias and Blair (2022; JOP)	Changing Tides: Public Attitudes on Climate Migration
57	Bayram and Graham (2022; JOP)	Knowing How to Give
58	McGrath et al. (2022; CPS)	Parliament, People or Technocrats? Explaining Mass Public Preferences on Delegation of Policymaking Authority
59	Bergquist et al. (2022; BJPS)	The Politics of Intersecting Crises: The Effect of the COVID-19 Pandemic on Climate Policy Preferences
60	Brutger and Guisinger (2022; JEPS)	Labor Market Volatility, Gender, and Trade Preferences
61	Carnegie and Gaikwad (2022; WP)	Public Opinion on Geopolitics and Trade Theory and Evidence
62	Duch and Gimeno (2022; CPS)	Collective Decision-Making and the Economic Vote
63	Frederiksen (2022; APSR)	Does Competence Make Citizens Tolerate Undemocratic Behavior?
64	Jurado et al. (2022; IO)	Brexit Dilemmas: Shaping Postwithdrawal Relations with a Leaving State
65	Krishnarajan and Jensen (2022; BJPS)	When Is A Pledge A Pledge?
66	Madsen et al. (2022; APSR)	Sovereignty, Substance, and Public Support for European Courts' Human Rights Rulings
67	Magni (2022; AJPS)	Boundaries of Solidarity: Immigrants, Economic Contributions, and Welfare Attitudes
68	Magni and Reynolds (2022; PB)	The Persistence of Prejudice: Voters Strongly Penalize Candidates with HIV
69	Manekin and Mitts (2022; APSR)	Effective for Whom? Ethnic Identity and Nonviolent Resistance
70	Rehmert (2022; PB)	Party Elites' Preferences in Candidates: Evidence from a Conjoint Experiment
71	Saha and Weeks (2022; PB)	Ambitious Women: Gender and Voter Perceptions of Candidate Ambition
72	Shandler et al. (2022; BJPS)	Cyber Terrorism and Public Support for Retaliation - A Multi-Country Survey Experiment
73	Simonsen and Bonikowski (2022; CPS)	Moralizing Immigration: Political Framing, Moral Conviction, and Polarization in the United States and Denmark
74	Weinberg (2022; CPS)	Feelings of Trust, Distrust and Risky Decision-Making in Political Office
75	Williams et al. (2022; APSR)	The Competing Influence of Policy Content and Political Cues
76	Williamson et al. (2022; BJPS)	Preaching Politics: How Politicization Undermines Religious Authority in the Middle East
77	Xu et al. (2022; JOP)	Information Control and Public Support for Social Credit Systems in China

Table OA-7: Multi-site Field Experiments

N	Author (Year; Journal)	Title
1	Green et al. (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 et al. (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
6	Baldwin (2013; AJPS)	Why Vote with the Chief? Political Connections and Public Goods Provision in Zambia
7	Broockman (2013; AJPS)	Black Politicians Are More Intrinsically Motivated to Advance Blacks' Interests
8	Findley et al. (2013; IO)	Using Field Experiments in International Relations: A Randomized Study of Anonymous Incorporation
9	Panagopoulos (2013; JOP)	Extrinsic Rewards, Intrinsic Motivation and Voting
10	Gift and Gift (2015; PB)	Does Politics Influence Hiring? Evidence from a Randomized Experiment
11	Nickerson (2015; JOP)	Do Voter Registration Drives Increase Participation? For Whom and When?
12	Nyhan and Reifler (2015; AJPS)	The Effect of Fact-Checking on Elites: A Field Experiment on US State Legislators
13	White et al. (2015; APSR)	What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials
14	Valenzuela and Michelson (2016; APSR)	Turnout, Status, and Identity: Mobilizing Latinos to Vote with Group Appeals
15	Broockman and Butler (2017; AJPS)	The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication
16	Findley et al. (2017; JOP)	External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation
17	Rooij and Green (2017; PB)	Radio Public Service Announcements and Voter Participation Among Native Americans
18	Grossman and Michelitch (2018; APSR)	Information Dissemination, Competitive Pressure, and Politician Performance between Elections
19	Kalla and Broockman (2018; APSR)	The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments
20	Kalla and Broockman (2020; APSR)	Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments
21	Linardi and Rudra (2020; CPS)	Globalization and Willingness to Support the Poor in Developing Countries: An Experiment in India
22	Persson et al. (2020; JEPS)	Does Deliberative Education Increase Civic Competence? Results from a Field Experiment
23	Choi et al. (2021; JEPS)	Linguistic Assimilation Does Not Reduce Discrimination Against Immigrants: Evidence from Germany
24	Harris et al. (2021; JOP)	Electoral Administration in Fledgling Democracies: Experimental Evidence from Kenya
25	Magni and Leon (2021; JEPS)	Women Want an Answer! Field Experiments on Elected Officials and Gender Bias
26	Moy (2021; JEPS)	Can Social Pressure Foster Responsiveness? An Open Records Field Experiment with Mayoral Offices
27	Bennion and Nickerson (2022; PB)	Decreasing Hurdles and Increasing Registration Rates for College Students
28	Goerger et al. (2022; JEPS)	Which Police Departments Want Reform? Barriers to Evidence-Based Policymaking
29	Lieberman and Zhou (2022; JEPS)	Self-Efficacy and Citizen Engagement in Development: Experimental Evidence from Tanzania

Table OA-8: Multi-site Lab Experiments

N	Author (Year; Journal)	Title
1	Aragones and Palfrey (2004; APSR)	The effect of candidate quality on electoral equilibrium: An experimental study
2	Wilking (2011; PB)	The Portability of Electoral Procedural Fairness: Evidence from Experimental Studies in China and the United States
3	Enos and Gidron (2016; JOP)	Intergroup Behavioral Strategies as Contextually Determined: Experimental Evidence from Israel
4	Vincent et al. (2016; JEPS)	The Electoral Sweet Spot in the Lab
5	Fournier et al. (2020; APSR)	Negativity Biases and Political Ideology: A Comparative Test across 17 Countries
6	Blais and Vallve (2021; PB)	Conformity and Individuals' Response to Information About Aggregate Turnout
7	de la Cuesta et al. (2022; JOP)	Owning It: Accountability and Citizens' Ownership over Oil, Aid, and Taxes

D.1.2 Observational Studies

For observational studies, we review 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, a two-step approach combining GPT-labeling and experts-manual-verification. In the first step, 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. In this verification step, we coded a total of 22 articles as a multi-site out of the 112 articles reviewed. Again, importantly, all studies we review below are manually verified to be multi-site observational studies. This means that the number of multi-site observational studies we report is likely the lower bound of the true number of multi-site observational studies that exist during the time frame we examine.

Table OA-9 displays the list of multi-site observational studies separated by the identification strategies: Difference-in-Difference, Regression Discontinuity, Instrumental Variables, and Natural Experiment, respectively. Note that articles may be listed more than once if multiple types of identification strategies were implemented (e.g., Grossman et al. (2017) use both instrumental variable and difference-in-difference approach).

Table OA-9: Multi-site Observational Studies

N	Author (Year; Journal)	Title
	Difference-in-Differences	
1	Grossman et al. (2017; JOP)	Government Fragmentation and Public Goods Provision
2	Singh (2019; AJPS)	Compulsory Voting and Parties' Vote-Seeking Strategies
3	Ziller and Goodman (2020; JOP)	Local Government Efficiency and Anti-immigrant Violence
4	Iversen and Rehm (2022; CPS)	Information and Financialization: Credit Markets as a New Source of Inequality
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
	Regression Discontinuity Design	
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	Dunning and Nilekani (2013; APSR)	Ethnic Quotas and Political Mobilization: Caste, Parties, and Distribution in Indian Village Councils
3	Folke et al. (2016; APSR)	The Primary Effect: Preference Votes and Political Promotions
4	Eggers et al. (2018; AJPS)	Regression Discontinuity Designs Based on Population Thresholds: Pitfalls and Solutions
5	Cavaille and Marshall (2019; APSR)	Education and Anti-Immigration Attitudes: Evidence from Compulsory Schooling Reforms across Western Europe
6	Velez and Newman (2019; AJPS)	Tuning In, Not Turning Out: Evaluating the Impact of Ethnic Television on Political Participation

Table OA-9: Multi-site Observational Studies

N	Author (Year; Journal)	Title
7	Holbein and Rangel (2020; JOP)	Does Voting Have Upstream and Downstream Consequences? Regression Discontinuity Tests of the Transformative Voting Hypothesis
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
10	Iversen and Rehm (2022; CPS)	Information and Financialization: Credit Markets as a New Source of Inequality
11	Olson and Stone (2022; PB)	The Incumbency Advantage in Judicial Elections: Evidence from Partisan Trial Court Elections in Six US States
12	Rau (2022; CPS)	Partisanship as Cause, Not Consequence, of Participation
13	Song (2022; QJPS)	The Rank Effect in Multimember District Elections
Instrumental Variables		
1	Daly (2014; BJPS)	State Strategies in Multi-Ethnic Territories: Explaining Variation in the Former Soviet Union and Eastern Bloc
2	Grossman et al. (2017; JOP)	Government Fragmentation and Public Goods Provision
Natural Experiment		
1	Jupille and Leblang (2007; IO)	Voting for change: Calculation, community, and Euro referendums
2	Malesky and Samphantharak (2008; QJPS)	Predictable Corruption and Firm Investment: Evidence from a Natural Experiment and Survey of Cambodian Entrepreneurs
3	Dassonneville et al. (2019; PB)	Compulsory Voting Rules, Reluctant Voters and Ideological Proximity Voting
4	Singh (2019; AJPS)	Compulsory Voting and Parties' Vote-Seeking Strategies
5	Bateson and Weintraub (2022; JOP)	The 2016 Election and America's Standing Abroad: Quasi-Experimental Evidence of a Trump Effect
6	Iversen and Rehm (2022; CPS)	Information and Financialization: Credit Markets as a New Source of Inequality

D.1.3 Other Empirical Methods

To assess the increased popularity of multi-site causal studies, we also counted published articles applying other widely-used empirical methods: conjoint experiment, text analysis, instrumental variables (IV), regression discontinuity design (RDD), and difference-in-difference (DID). For each empirical method, we used the following keywords: "conjoint analysis" or "conjoint experiments" for conjoint analyses, "text as data" and "text analysis" for text analyses, "difference-in-difference" or "two-way fixed effects" for difference-in-difference (DID), "instrumental variable" for instrumental variable (IV), and "regression discontinuity" for regression discontinuity (RDD). For text analyses, we also included the articles published in the above-mentioned top journals that cite Grimmer and Stewart (2013). See Figure 1 of the main paper.

D.2 Descriptive Analyses of Multi-Site Studies

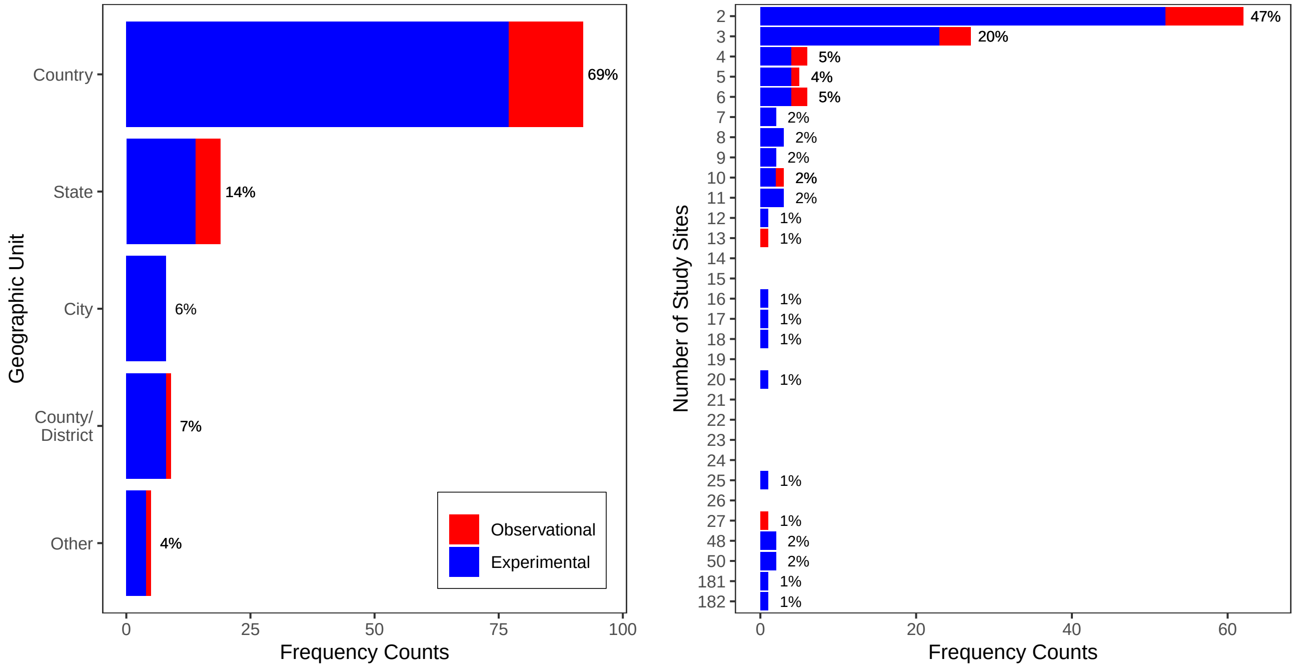
To further assess the current sampling approach in multi-site studies, we hired two independent researchers to review the 133 verified multi-site experimental (111) and observational (22) studies and code the following information:

1. The geographic unit as well as total number of study sites;
2. Use of random sampling in selecting study sites;
3. Use of purposive sampling in selecting study sites; and
4. Site-level variables considered when diversifying the site selection

In terms of the current practice of site selection, random sampling of sites is extremely rare. Among all 133 multi-site studies we review, we found only 2 papers using random sampling. Instead of using random sampling, about 80% of multi-site studies rely on purposive sampling and select diverse sites such that study sites cover heterogeneous contextual factors.

We next examine geographic unit of study sites and how many sites are involved in multi-site studies. First, as shown in Figure OA-5-(a), large majority of the multi-site studies (69%) are implemented at the country-level, which reflects the increased popularity of multi-country survey experiments. However, it is also important to emphasize that multi-site causal studies are also used within a country, across states, cities, counties, districts, and so on. SPS can be used for any geographic unit of study sites and is equally effective for selecting multiple countries and multiple states/cities/counties/districts within a country.

Figure OA-5-(b) shows the distribution of the number of study sites researchers select in each paper. The median number of study sites is 3 and the 80th percentile is 6.6. In general, the number of study sites is small in political science. This is one of the main reasons why random sampling of sites is often infeasible and ineffective in practice. We emphasize that SPS is designed specifically for this small-sample setting.



(a) Geographic Unit of Study Sites

(b) Number of Study Sites

Figure OA-5: Breakdown of Multi-Site Studies

E Formal Results

E.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 4.2.2, 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), which is defined as follows.

$$\text{MSE} := \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left\{\left(\theta_k - \widehat{\theta}_k^W\right)^2\right\}$$

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.

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

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 $d_j := \widehat{\theta}_j - \theta_j$. 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 $k \in \{1, \ldots, N\}$,

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

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., 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 for non-selected site k. Importantly, the following decomposition holds for any weighted average estimators.

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

where the first line follows from the definition of a weighted average estimator, the second from the definition of d described above, and the third from rearrangement of terms, and the final line from the definition of η_k.

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

$$\begin{aligned}
||F_k(\mathbf{W})||_2^2 \; := \; & \left(f(g(\mathbf{X}_k)) - \sum_{j \in \mathcal{S}} W_{jk} f(g(\mathbf{X}_j)) \right)^2 \\
\leq \; & \left(\rho \times \sum_{j \in \mathcal{S}} W_{jk} ||g(\mathbf{X}_k) - g(\mathbf{X}_j)||_2 \right)^2
\end{aligned}$$

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

where we use $[N] \coloneqq \{1, \ldots, N\}$. The first line follows from the definition and the second from the property of the Lipschitz function $f(\cdot)$ with Lipschitz constant $\rho \geq 0$. The third follows from $\sum_{j' \in \mathcal{S}} W_{j'k}$ being the weighted average, the fourth 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

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

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,

$$\mathbb{E}\left\{(\sum_{j\in\mathcal{S}} W_{jk}d_j)^2\right\} = \sum_{j\in\mathcal{S}} W_{jk}^2\mathbb{E}(d_j^2) = \sum_{j\in\mathcal{S}} W_{jk}^2 \mathrm{Var}(\widehat{\theta}_j) \leq \max_{j'\in[N]} \mathrm{Var}(\widehat{\theta}_{j'}) \sum_{j\in\mathcal{S}} W_{jk}^2$$

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, for non-selected site k,

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

where

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

Finally, we take the average of the MSE over all sites.

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N} \mathbb{E}\left\{\left(\theta_k - \widehat{\theta}_k^W\right)^2\right\} \\
=\ &\frac{1}{N}\sum_{k\in\mathcal{R}} \mathbb{E}\left\{\left(\theta_k - \widehat{\theta}_k^W\right)^2\right\} + \frac{1}{N}\sum_{j\in\mathcal{S}} \mathbb{E}\left\{\left(\theta_j - \widehat{\theta}_j\right)^2\right\}
\end{aligned}$$

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

where

$$\begin{aligned}
\lambda_1 \quad &:= \quad \frac{N - N_S}{N} \times L_g \times ||\beta||_2 \times \left(||\beta||_2 + 2\rho \times \max_{j', k} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2 + 4\overline{\eta} \right) \\
\lambda_2 \quad &:= \quad \frac{N - N_S}{N} \times \left(L_g \times \rho^2 \times \max_{k \in [N]} \frac{\max_{j' \in [N]} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2}{\min_{j' \in [N]} ||g(\mathbf{X}_k) - g(\mathbf{X}_{j'})||_2} + 4\rho\overline{\eta} \right) \\
\lambda_3 \quad &:= \quad \frac{N - N_S}{N} \times \left(\overline{\eta}^2 + \max_{j' \in [N]} \mathrm{Var}(\widehat{\theta}_{j'}) \right).
\end{aligned}$$

When we set $C = \frac{N_S}{N}(4\overline{\eta}^2 + \max_{j' \in [N]} \mathrm{Var}(\widehat{\theta}_{j'}))$, 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. □

E.2 Solving SPS Optimization Problem

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

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

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

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

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

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.

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

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

$$\min_{(\mathbf{S} \in \{0,1\},\ \mathbf{W}, \mathbf{Q}, \mathbf{Z})} \quad \lambda_1 \times \frac{1}{(N - N_S) L_g} \sum_{k=1}^{N} \sum_{\ell=1}^{L_g} Z_{k\ell}^2 \tag{OA.5}$$

$$+ \lambda_2 \times \frac{1}{(N - N_S)} \sum_{j=1}^{N} \sum_{k=1}^{N} Q_{jk} \frac{1}{L_g} \sum_{\ell=1}^{L_g} (g_\ell(\mathbf{X}_j) - g_\ell(\mathbf{X}_k))^2, \tag{OA.6}$$

$$+ \lambda_3 \times \frac{1}{N - N_S} \sum_{j=1}^{N} \sum_{k=1}^{N} Q_{jk}^2 \tag{OA.7}$$

$$\text{such that} \qquad \sum_{k=1}^{N} S_k = N_S, \quad \mathbf{W} \geq 0, \ \sum_{j=1}^{N} Q_{jk} = 1 - S_k, W_{jk} \leq 1 - S_k, \tag{OA.8}$$

$$Z_{k\ell} = (1 - S_k) g_\ell(\mathbf{X}_k) - \sum_{j=1}^{N} Q_{jk} g_\ell(\mathbf{X}_j) \tag{OA.9}$$

$$0 \leq Q_{jk} \leq S_j, \text{ and } \ W_{jk} - (1 - S_j) \leq Q_{jk} \leq W_{jk} \tag{OA.10}$$

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 and Gurobi) 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

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

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

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,

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

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

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

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

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

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

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 $\sum_j S_j W_{jk} = 1$ 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 $Q_{jk} = S_j W_{jk}$. This completes the proof. □

E.3 Extending SPS to Accommodate More Domain Knowledge

Researchers can also incorporate various other domain knowledge to SPS. None of the following extensions fundamentally change the theoretical properties of SPS, but it improves flexibility.

First, as we analyzed formally in equation (OA.1), users can incorporate not only $\mathbf{X}_k$ themselves but also any flexible functions of site-level variables $g(\mathbf{X}_k)$, e.g., interaction between GDP and population size and higher order terms like squared population size (Population^2), to capture nonlinearity in the data.

Second, researchers can also incorporate varying importance weights. For example, researchers can extend equations (OA.1) and (OA.2) as follows.

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

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

where VW_ℓ is the importance weight for the ℓth site-level variable. As in the standard synthetic control method, researchers might choose these weights based on predictive power.

Third, users can also ensure that selected sites are geographically diverse and distant enough from each other. For example, if users want to ensure the minimum distance (e.g., 20 km) between each selected site, researchers can add the following constraint to SPS. For each pair (j, k) with $S_j S_k = 1$, we add

$$\text{distance}_{jk} \quad \geq \quad \text{minimum distance.}$$

If users just want to encourage geographically diverse sites, they can add the following term to the objective function.

$$-\sum_{j=1}^{N}\sum_{k=1}^{N} S_j S_k \times \text{distance}_{jk}$$

Fourth, researchers can incorporate the budget constraint and differential costs of each site by adding the following to SPS.

$$\sum_{j=1}^{N} S_j \text{cost}_j \leq \text{Total Budget}, \tag{OA.11}$$

where cost_j captures the cost of conducting a causal study in site j. This will include different types of costs, such as the initial cost of setting up experiments, hiring local partners, recruting subjects, and so on. This calculation is often easier when considering multi-country survey experiments and online survey firms give researchers costs per subject in each country.

Fifth, researchers can also incorporate differential sample size in each site. In particular, differential sample size in each site affects the expected variance in each site and thus we can naturally incorporate it into equation (OA.3).

$$\lambda_3 \times \frac{1}{N - N_S} \sum_{j=1}^{N}\sum_{k=1}^{N} \frac{1}{n_j} S_j (1 - S_k) W_{jk}^2$$

where n_j is the sample size in site j. This is because the third term (equation (OA.3)) is related to the variance of the site-specific ATE estimator $\text{Var}(\widehat{\theta}_j)$, which is a function of $1/n_j$. See equation (OA.5). Differential sample size also naturally affects the cost in site j, which is incorporated in equation (OA.11).

E.4 SPS Estimators

E.4.1 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

$$\widehat{\theta}_{AS} \quad := \quad \frac{1}{N}\Big(\sum_{j \in \mathcal{S}} \widehat{\theta}_j + \sum_{k \in \mathcal{R}} \widehat{\theta}_k^W\Big) = \frac{1}{N}\Big(\sum_{j \in \mathcal{S}} \widehat{\theta}_j + \sum_{k \in \mathcal{R}}\sum_{j \in \mathcal{S}} \widehat{W}_{jk} \widehat{\theta}_j\Big)$$

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

where the first and second equalities follow from the definition of the SPS estimators, the third equality follows from from rearrangement of terms.

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

$$\widehat{\theta}_{AS} = \sum_{j\in\mathcal{S}}\widetilde{W}_j\widehat{\theta}_j \quad \text{where} \quad \sum_{j\in\mathcal{S}}\widetilde{W}_j = 1.$$

E.4.2 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.

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

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., $\mathbb{E}(\widehat{\tau}^2) \geq \tau^2$.

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

where

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

$\widehat{\sigma}_k^2$ 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 $\widehat{e}_k$.

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

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,

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

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

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

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. □

E.4.3 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.

E.5 Connections to and Differences from Meta-Analysis Estimators

The proposed SPS estimator is strongly connected to typical meta-analysis estimators. 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; Dunning *et al.*, 2019). 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. We discuss its relationship to SPS in Section 7.4.

F Simulation Study

In this section, we provide a simulation study to investigate the finite sample performance of SPS.

Simulation Design. We use data from Naumann *et al.* (2018) to mimic the real-world data generating process. In particular, we regress the site-specific ATEs in 15 European countries on 5 key site-level variables (GDP, Size of Immigration Population, Unemployment Rate, Mean Age, and General Support level for Immigrants) and squared terms of GDP and General Support level for Immigrants. We then use the estimated coefficients as the true coefficients for the simulation study. In this way, we can design simulation studies similar to the real-world application, while we can control several key parameters to investigate the properties of SPS.

We change two key parameters. (1) The number of study sites we select, $N_S \in \{3, 6, 9\}$. These numbers cover from small to moderate and large multi-site studies in political science (see Appendix D). (2) The number of covariates we include in SPS, $L \in \{3, 4, 5, 7, 9, 12, 15, 20, 25, 30\}$. Importantly, the number of relevant site-level variables is 5. By changing the number of variables we include in SPS, we can explore both settings where users miss relevant variables, i.e., unobserved site-level variables (when $L \in \{3, 4\}$) and settings where users include irrelevant variable (when $L > 5$).

We then compute the root mean squared error (RMSE) of the Average-Site ATE estimator. We compare SPS against random sampling of sites, which is used here as a theoretical benchmark, even though random sampling of sites is often infeasible in practice.

Results. Figure OA-6 shows the results. First, when the number of study sites is within a range of political science studies, SPS achieves lower RMSE than that of random sampling, denoted by the red dotted lines in each panel. As known in the literature, when the number of study sites is small, random sampling of sites has too large standard errors and is unstable, while it is unbiased. This gain by SPS is especially large when the number of study sites is small ($N_S = 3$). Second, RMSE of SPS is the lowest when we include all relevant variables but we do not include irrelevant variable ($L = 5$). RMSE increases when we include irrelevant variables (when $L > 5$), while its increase is relatively moderate. This is because SPS decreases the diversity in key site-level variables to improve the diversity in irrelevant variables. Therefore, we recommend against a kitchen sink approach of including too many irrelevant variables. Finally, SPS has relatively low RMSE even when users cannot include all relevant variables and there are unmeasured moderators (when $L < 5$). Even when users miss some key site level variables, RMSE of SPS is still lower than RMSE of random sampling approach. The SPS algorithm can reduce RMSE further if users can include more relevant site-level variables, but unobserved moderators do not invalidate the use of SPS.

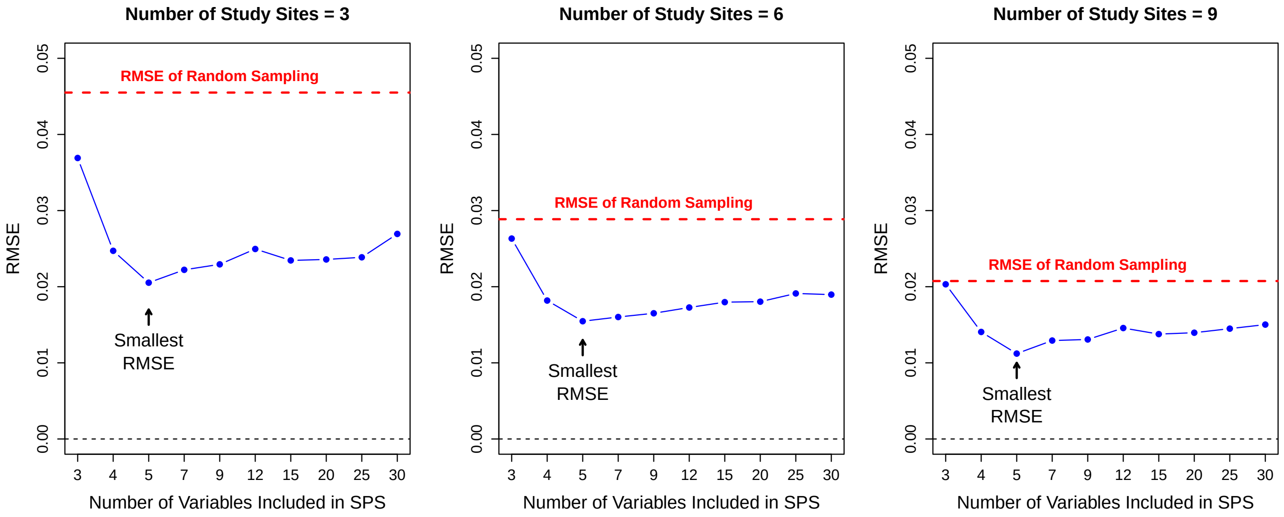


Figure OA-6: **Simulation Results for SPS.** *Note*: Blue lines represent RMSE of SPS, while the red dotted lines represent RMSE of random sampling.

References

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

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.

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

Bisbee, J., Dehejia, R., Pop-Eleches, C., and Samii, C. (2017). Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect. *Journal of Labor Economics* **35**, S1, S99–S147.

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.

de Benedictis-Kessner, J., Lee, D. D. I., Velez, Y., and Warshaw, C. (2022). Local representation in the united states: A new comprehensive dataset of elections .

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

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.

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.

Freedom House (2022). Lithuania: Nations in transit 2022 country report.

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

Gift, K. and Gift, T. (2015). Does Politics Influence Hiring? Evidence from A Randomized Experiment. *Political Behavior* **37**, 653–675.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* **21**, 3, 267–297.

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.

Lyall, J., Shiraito, Y., and Imai, K. (2015). Coethnic bias and wartime informing. *The Journal of Politics* **77**, 3, 833–848.

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.

U.S. Department of State (2022a). 2022 country reports on human rights practices: Bolivia.

U.S. Department of State (2022b). 2022 country reports on human rights practices: Kenya.

U.S. Department of State (2022c). 2022 country reports on human rights practices: Lithuania.

Voeten, E., Strezhnev, A., and Bailey, M. (2009). United Nations General Assembly Voting Data.