

# The Ethics of Electoral Experimentation: Design-Based Recommendations

Tara Slough\*

September 9, 2022

## Abstract

While experiments on elections represent a popular tool in social science, the possibility that experimental interventions could affect who wins office remains a central ethical concern. I formally characterize electoral experimental designs to derive an upper bound on aggregate electoral impact under different assumptions about interference. I then introduce a decision rule based on comparison of this bound to predicted election outcomes to determine whether an experiment should be implemented. I demonstrate that existing experiments vary substantially in their (*ex-ante*) risk of changing aggregate electoral outcomes. Researchers can mitigate the possibility of affecting aggregate outcomes by reducing the saturation of treatment or focusing experiments in districts where treated voters are unlikely to be pivotal. These conditions identify novel trade-offs between adhering to ethical commitments and the statistical power and external validity of electoral experiments. More broadly, I show that deliberate research design can address some ethical concerns with experiments.

---

\* Assistant Professor, New York University, [tara.slough@nyu.edu](mailto:tara.slough@nyu.edu). I am indebted to Jiawei Fu and Kevin Rubio for expert research assistance. I thank Eric Arias, Graeme Blair, Alex Coppock, Sandy Gordon, Saad Gulzar, Macartan Humphreys, Kimuli Kasara, Dimitri Landa, Eddy Malesky, John Marshall, Lucy Martin, Kevin Munger, Gareth Nellis, Franklin Oduro, Melissa Schwartzberg, Lauren Young, attendees of APSA 2019, and students in NYU's graduate Scope and Methods class for generous feedback. This project is supported in part by an NSF Graduate Research Fellowship, DGE-11-44155.

# 1 Introduction

Experiments on real elections represent a popular tool in studies of elections, political behavior, and political accountability. While the use of experiments on elections dates back nearly a century to Gosnell (1926), the scale, sophistication, and frequency of electoral experiments has increased precipitously since the late 1990s. A central ethical concern in the study of electoral experiments is that by manipulating characteristics of campaigns, candidates, or voter information, researchers may also be changing aggregate election outcomes.

Two notable changes in experimental studies of elections over the past two decades influence these ethical considerations. First, researchers now work in contexts with greater variation in institutions and voting behavior than early studies of elections in US college towns.<sup>1</sup> Figure 1 draws on a original dataset of all pre-registered electoral experiments in the American Economic Association (AEA) and Evidence in Governance and Politics (EGAP) experimental registries through 2020.<sup>2</sup> It shows that the modal pre-registered experiment is conducted outside the US (57% versus 43%). Second, the scale of electoral interventions, measured in terms of the number of treated voters, has increased precipitously. In addition to researchers, campaigns and technology companies now implement massive experimental interventions in elections (i.e., Pons, 2018; Bond et al., 2012).

Turning to the content of these experiments, Figure 1 shows the frequency of different classes of interventions. Many of these interventions present limited risks of *individual* harm to subjects—generally registered voters. For example, most mobilization or “get out the vote” interventions consist of a phone call, pamphlet, or visit from a canvasser. Most pre-electoral information interventions consist of pamphlets, meetings, or media broadcasts. However, even when these interventions present minimal risk of individual harms, they can produce substantial risk of downstream *social* harms to subjects and non-subjects alike. In elections, the most obvious mechanism for these social harms is changes in aggregate electoral outcomes, or who wins office. In other words,

---

<sup>1</sup>Early (pre-2000) experiments from Gosnell (1926) to Gerber and Green (2000) occurred in local elections in jurisdictions where researchers worked, namely Chicago, Ann Arbor, and New Haven.

<sup>2</sup>See Appendix A3 for a description of the data.

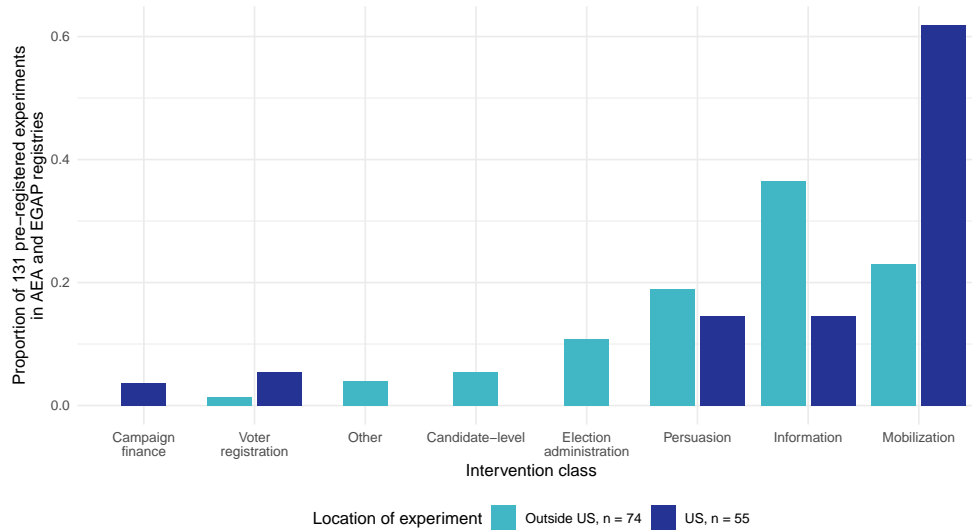


Figure 1: Types of experiments in pre-registered experiments in and outside the US. See Table A2 for clarification of the intervention classes.

by inducing more individuals to turn out or change their votes, experimental elections have the potential to flip election outcomes.

This risk of downstream social harms is well-known in existing literature on experimental ethics. Teele (2013: p. 117) writes that a “more thoroughgoing evaluation of the downstream and community-level risks that stem from field experiments must guide all research if it is to be ethical.” Phillips (2021: p. 281) emphasizes that: “the process-related downstream effects of these interventions can create winners and losers and harm individuals and groups. They can also harm entire communities.” In the context of elections, McDermott and Hatemi (2020: p. 30015) state that “shifting the actual outcome of an election has real effects on local and national society.” McDermott and Hatemi (2020), Gubler and Selway (2016), and Zimmerman (2016) further elaborate particular concerns about the potential for disparate welfare impacts of these downstream consequences across subjects, non-subjects, or communities in the context of elections. These potential downstream consequences of electoral interventions can be quite difficult to predict *ex-ante*. Carlson (2020: p. 92) argues that “even expert researchers and their local partners can fail to correctly predict the outcome of their interventions,” which can imply that they “do not have sufficient information to anticipate harm.” Indeed, Baele (2013: p. 28) suggests that “artificial interventions in

tension-ridden political episodes such as elections may provoke unpredictable chain reactions.”

These arguments have led to ethical guidance that electoral experiments should be designed such that they are sufficiently unlikely to change aggregate election outcomes. Recently-adopted American Political Science Association (APSA) “Principles and Guidance for Human Subjects Research” echo this concern, stating that interventions are of “minimal social risk if they are not done at a scale liable to alter electoral outcomes” (American Political Science Association, 2020: p. 15). To this point, Desposato (2016: p. 282) advocates “treading lightly” in an effort to “minimiz[e] the aggregate social costs and minimiz[e] the amount of deception.” He further argues that “if running an election field experiment, treading lightly implies choosing an election where polling and history suggest that your treatment, in the worst-case scenario does not affect the outcome” (Ibid).

To what extent do researchers adhere to this guidance when designing electoral experiments? Some authors report undertaking these considerations in published research. For example, Dunning et al. (2019: p. 52) write that the authors of seven coordinated experiments on elections and accountability “elaborated research designs to ensure to the maximum extent possible that our studies would not affect aggregate election outcomes.” However, the assembled pre-registered experiments from the AEA and EGAP registries tell a different story. Of the 129 experiments in Figure 1, just two discuss the possibility of changing aggregate outcomes among their written *ex-ante* design considerations (Table A3). This article proposes a design-based approach to the *ex-ante* consideration of how experimental interventions could affect aggregate election outcomes.

Minimizing the possibility of changing aggregate electoral outcomes requires two departures from standard practice in the design and analysis of experiments. First, consideration of election outcomes requires aggregation to the level of the *district*. The district is rarely the level at which treatment is assigned or outcomes are analyzed. The frequent omission of information about the relationship between the electoral district and experimental units (of assignment or outcome measurement) makes it difficult to estimate *ex-post* the saturation of an intervention in the relevant electorate in many existing studies.

Second, while experiments are powerful tools for estimating various forms of *average* causal effects, the relevant ethical consideration is whether an electoral experiment changes *any* individual election outcome, defined here in terms of who wins office. Yet, such individual (district-level) effects are unobservable due to the fundamental problem of causal inference. Furthermore, any *ex-post* attempt to assess electoral impact must acknowledge that the possible consequences of an electoral intervention are set into motion when the experiment goes into the field. For this reason, I suggest that the relevant course of action is to consider the possible impact of an experimental intervention *ex-ante*. In this sense, I examine how to design experiments that are unlikely to change who wins office via the random assignment of treatment.

I propose a framework for bounding the maximum aggregate electoral impact of an electoral experiment *ex-ante*. I focus on the design choices made by researchers, namely the selection of districts (races) in which to implement an intervention and the saturation of an intervention within that electorate. With these design choices, I allow for maximum voter agency in response to an electoral intervention through the invocation of “extreme value bounds” introduced by Manski (2003). Combined with assumptions about interference between voters, this framework allows for the calculation of an experiment’s maximum aggregate electoral impact in a district. The relevant determination of whether an intervention should be attempted rests on how this impact compares to predicted electoral outcomes in a district. I propose a decision rule that can be implemented to determine whether or not to run an experimental intervention. I complement the analysis with an R package for easy calculation of these bounds and the decision rule.

This analysis identifies a set of experimental design decisions that researchers can make to minimize the possibility of changing election outcomes. They can reduce the saturation of treatment in a district by (1) treating fewer voters or (2) intervening in larger districts. Further, they can avoid manipulating interventions in (3) close or unpredictable contests or (4) PR contests. These design principles suggest trade-offs between ethical considerations and learning from electoral experiments.

While the analysis is agnostic with respect to voter responses to an experimental intervention, I

show that some assumption restricting interference between voters is necessary for an experiment to ever pass the proposed decision rule.<sup>3</sup> I derive bounds on the maximum electoral impact under the stable unit treatment value assumption (SUTVA) as well as weaker and stronger assumptions about interference (spillovers). Because these assumptions must be invoked *ex-ante*, more careful consideration of possible general equilibrium effects should be central to ethical considerations.

In the discussion, I reflect on the ethical merits of guidance to avoid changing electoral outcomes. I discuss more and less restrictive arguments about experimentation in real elections. I argue that the guidance to avoid changing electoral outcomes is consistent with the principle of beneficence and represents a useful default when designing or evaluating electoral experiments.

This paper makes three contributions. First, it develops tools to guide researchers considering prospective interventions on elections, as well as consumers of this research. Second, I identify a set of trade-offs inherent to the design of electoral experiments that emerge from considerations of aggregate electoral impact. Characterization of these trade-offs allows for a richer discussion about the merits and limitations of experiments on elections as a research design. Finally, I advance the first general framework to incorporate ethical concerns across a range of experimental designs in a common setting (elections). I outline how this basic structure can be extended to inform the design of field experiments in other domains where changing aggregate social or political outcomes represents a salient ethical concern.

## **2 Experiments and their counterfactuals**

In order to consider how much an experimental intervention might affect an electoral outcome, experimentalists should ask “what would have happened absent the experimental intervention?” The answer to this question generally depends on whether researchers are implementing their own intervention or whether they are randomizing an intervention that a partner would have implemented regardless. In the context of elections, these partners are typically political parties, NGOs, or candidates. The American Political Science Association (2020) and Hyde and Nickerson (2016) sug-

---

<sup>3</sup>See Beerbohm, Davis, and Kern (2020) and Michelson (2016) for discussion of the moral status of specific types of treatments classified in Figure 1.

gest that the ethical considerations with these partnerships depart from those of researcher-initiated and implemented interventions, advocating a lesser level of researcher responsibility. Importantly, these partners are typically motivated by some type of normative goal such as winning elections or improving some aspect of governance. In these collaborations, researchers often study the effects of interventions that are explicitly aimed at changing electoral processes or outcomes.

I contend that the relevant consideration in cases of partnerships is how aggregate outcomes may be changed by *random assignment* of treatment. Partners' pre-existing electoral goals likely guide how partners target interventions outside of an experiment. For example, an anti-corruption partner organization may prefer to target districts where corruption is worse or competitive districts where less corrupt candidates stand a better shot at winning. By randomly assigning the intervention, however, researchers may move the intervention away from the races in which it stands the best shot at achieving a partner's stated goal to reduce corruption. In this setting, the use of random assignment to assign a partner's well-intentioned and effective intervention may *reduce* welfare of subjects and non-subjects in the electorate relative to its non-experimental allocation. Given these potential harms, I argue that in collaborations, researchers are responsible for how the random assignment of the intervention—in the service of research—changes the allocation of the treatment. In this sense, when collaborating with partners, researchers may be justified in studying interventions intended to change electoral outcomes, but these impacts should not be induced by the research component, specifically random assignment.

The preceding discussion suggests two possible counterfactuals to electoral experiments, as a function of the involvement of a partner. These cases are summarized in Table 1. Case #1 describes electoral experiments in which a researcher designs and implements an intervention that would not have otherwise occurred. Case #2 considers the change in the allocation of a partner's intervention to accommodate the random assignment of the intervention. As such, the counterfactual is the partner's allocation of the intervention as opposed to no intervention.

Because of the aggregation of votes at the district level, changes between an experimental and counterfactual allocation of an intervention take three forms, as depicted in Table 2. First and

Case	Actors		Experiment	Counterfactual (absent experiment)	Examples
	Researcher	Partner			
1	✓		<i>Researcher designs, implements experimental intervention.</i> (Note: An partner may participate in or endorse the experiment, but the researcher causes the intervention to occur. Such interventions are often, but not necessarily, funded through the researcher.)	<i>No intervention occurs.</i>	Gerber and Green (2000); Metaketa-I experiments documented in Dunning et al. (2019)
2	✓	✓	<i>Researcher randomizes a partner-funded and implemented intervention.</i>	<i>Partner funds and implements intervention without randomizing allocation of treatment, possibly with less data collection.</i>	Bond et al. (2012), Kendall, Nannicini, and Trebbi (2015), Pons (2018), López-Moctezuma et al. (2021)

Table 1: Classification of experiments and their counterfactuals by the actors involved in experimental design and implementation.

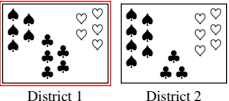
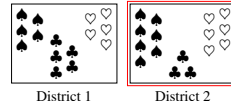
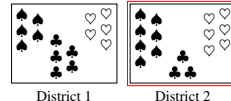
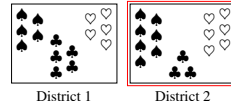
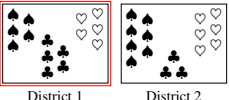
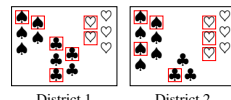
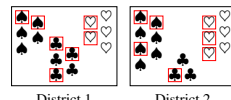
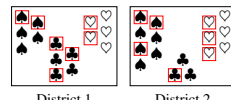
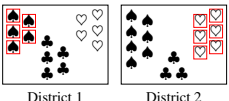
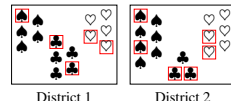
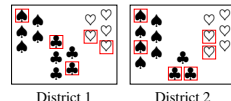
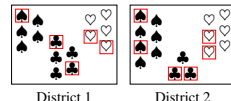
	Counterfactual allocation		Experimental allocation		Description
1					Intervention is implemented in <b>different districts</b> under experimental allocation.
2					Intervention is implemented at <b>different saturation</b> under experimental allocation.
3					Intervention is targeted to <b>different voters</b> under experimental allocation.

Table 2: Each symbol represents a voter. The red boxes indicate treatment assignment.

closest to Case #1, the intervention may be implemented in districts that where it would otherwise not have been implemented in the non-experimental regime. Second and generalizing this point, there may be a change in the proportion of a district that is treated (differential saturation). Finally, it may be the case that different voters in a district are treated under an experimental allocation of treatment. Thus, even holding constant the number of treated voters in a district, the types (and thus voting behavior) of treated voters may change when researchers randomly assign treatment.

Ethical concerns about an experiment changing aggregate electoral outcomes should focus on the difference in treatment allocation between the experiment and its counterfactual. Because of the aggregation of votes at the district level in the context of elections, subjects are not simply “interchangeable” when the allocation of treatment is changed. As such, within this framework,



the guideline that “studies of interventions by third parties do not usually invoke [the principle of not impacting political outcomes]” put forth by American Political Science Association (2020: p. 14) is insufficient. Attention to the contrast between experimental and counterfactual allocation of an intervention arguably formalizes Hyde and Nickerson’s (2016) concept of an intermediate level of scrutiny for experiments conducted with partners that is lower than the level of scrutiny afforded to experiments conducted without a partner. In sum, collaboration with a partner does not absolve a researcher from concerns of aggregate electoral impact, though it restricts focus to a more specific set of considerations.

Reporting this counterfactual allocation of treatment is not yet standard practice, making the risks of experiments conducted in partnerships challenging to assess. Two questions are critical. First, how did the experiment change the allocation of the treatment? Second, how similar was the experimental treatment to the intervention that would have been implemented absent the experiment? When researchers have a role in shaping partners’ interventions, they should justify whether Case #1 or Case #2 better describes the character of the partnership and proceed accordingly.

## **2.1 The Ethical Objective**

I assume researchers’ ethical objective is to avoid changing who ultimately wins office relative to what would have happened absent the experimental allocation of an intervention. As such, researchers would ideally minimize the probability that their interventions change the *ex-post* distribution of seats or offices. In so doing, I assume that the primary electoral consequences on policymaking or governance occur because candidate *A* wins office, not because candidate *A* won office with 60 percent instead of 51 percent of the vote. Effects based on vote share of the winner can be considered “mandate effects.”

Some literatures suggest that the assumption of no mandate effects may be too stark. Politicians may condition resource allocation on expressed political support, if distributional targets are informed by election outcomes (Lindbeck and Weibull, 1987; Catalinac, Bueno de Mesquita, and Smith, 2020). Alternatively, in uncompetitive electoral autocracies, vote shares – particularly in the form of supermajorities – are argued to signal regime strength (e.g., Simpser, 2013). Changing

vote shares in such a context may have consequences for governance or regime stability.

While this paper abstracts from these types of mandate effects, it also provides the tools to rigorously develop such considerations. The calculation of aggregate electoral impact is not affected by the specific ethical objective. The decision rule, however, does depend on how the objective is specified. The exceptions outlined above can be formalized as alternative decision rules in cases where mandate effects are particularly likely or concerning. Thus, this article provides tools that can be developed in the design of experiments in settings where the specific objective – in terms of election outcomes – varies.

The framework starts from the observation that we do not know precisely what an election outcome would be in the absence of an experimental manipulation. This limits our ability to design an experiment to minimize the probability that their interventions change the *ex-post* distribution of office holders or ballot outcomes. As such, this paper advocates the estimation of conservative bounds on the *ex-ante* possible change in vote share. These bounds can be calculated analytically. I then develop a decision rule that compares these bounds to the predicted closeness of an election in order to minimize the risks of altering electoral outcomes. Through these steps, I argue that research can be designed or avoided as to minimize these risks. By reporting these quantities in grant applications, pre-analysis plans, and ultimately research outputs, researchers can transparently justify their design choices.

### **3 Formalizing the Design of Electoral Experiments**

I proceed to construct bounds with three sets of considerations: design decisions made by researchers; researcher assumptions about which voters' treated potential outcomes are revealed by the intervention; and a minimal model of voter behavior that is sufficiently general to encompass many types of electoral experiment. Collectively, these considerations allow researchers to calculate a conservative bound on the extent to which an experiment could change election outcomes.

### 3.1 Research Design Decisions

I first consider the components of the research design controlled by the researcher, potentially in collaboration with a partner (as in Case #2). The researcher makes three design decisions. First, she controls the set of districts,  $D$ , in which to experimentally manipulate an intervention. Indexing electoral districts by  $d \in D$ , the number of registered voters in each district is denoted  $n_d$ .

Second, researchers define the clustering of subjects within a district. I assume that voters in district  $d$ , indexed by  $j \in \{1, \dots, n_d\}$  are partitioned into  $C$  exhaustive and mutually exclusive clusters. I index clusters by  $c \in C$  and denote the number of voters in each cluster by  $n_c$ , such that  $\sum_c n_c = n_d$ . There is always a cluster, even when treatments are not cluster-assigned. Individual-level (voter-level) randomization can be accommodated by assuming  $n_c = 1 \forall c$ . Similarly, district-level clustering can be accommodated by assuming  $n_c = n_d$ . In practice, researchers typically assign electoral interventions to individuals or precincts (generally below the district level).

Finally, researchers decide the allocation of treatment within a district. Consider two states of the world,  $E \in \{e, \neg e\}$ , where  $e$  indicates an experiment and  $\neg e$  indicates no experiment. These states represent the counterfactual pairs described in Table 1. Our main potential outcome of interest,  $\pi(E)$  is whether an individual voter is assigned to receive a treatment. In the experiment, allocation occurs via random assignment. Absent an experiment, the treatment could be assigned by any allocation mechanism. This notation allows for characterization of four principal strata, described in Table 3. Note that by asserting the possibility of four strata, I allow for cases in which a researcher's partner would assign any proportion of the electorate (including all or none) to the intervention in the absence of the experiment. I use the notation  $S_{11}^{cd}$ ,  $S_{10}^{cd}$ ,  $S_{01}^{cd}$ , and  $S_{00}^{cd}$  to denote the set of voters belonging to each stratum in each cluster and district. The cases defined in Table 1 place assumptions on the relevant strata. Where the counterfactual is no intervention (Case #1), strata where  $\pi(\neg e) = 1$  must be empty.

Stratum		Intervention		Assumptions	
Set	Name	$\pi(e)$	$\pi(\neg e)$	Case 1	Case 2
$S_{11}^{cd}$	Always assigned	1	1	$ S_{11}^{cd}  = 0$	$ S_{11}^{cd}  \geq 0$
$S_{10}^{cd}$	If-experiment assigned	1	0	$ S_{10}^{cd}  > 0$	$ S_{10}^{cd}  \geq 0$
$S_{01}^{cd}$	If non-experiment assigned	0	1	$ S_{01}^{cd}  = 0$	$ S_{01}^{cd}  \geq 0$
$S_{00}^{cd}$	Never assigned	0	0	$ S_{00}^{cd}  > 0$	$ S_{00}^{cd}  \geq 0$

Table 3: Principal strata. Each individual (registered voter) belongs to exactly one stratum. The cases refer to those described in Table 1. The  $|\cdot|$  notation refers to the cardinality of each set, or the number of voters in each stratum in cluster  $c$  in district  $d$ .

With this notation, I proceed by characterizing the proportion of a district’s electorate that is assigned or not assigned to the treatment *because* of the experiment. From Table 3, the relevant strata are  $S_{10}^{cd}$ —individuals exposed to the treatment because it is assigned experimentally—and  $S_{01}^{cd}$ —individuals not exposed to the treatment because is assigned experimentally. The proportion of the electorate in a district that is exposed (resp. not exposed) to an intervention due to the experiment, heretofore the *experimental saturation*,  $\mathcal{S}_d$  can thus be written:

$$\mathcal{S}_d = \frac{\sum_{c \in d} |S_{10}^{cd} \cup S_{01}^{cd}|}{n_d} \quad (1)$$

In the context of electoral interventions that would not occur absent the experiment (Case 1), the interpretation of  $\mathcal{S}_d$  is natural: it represents the proportion of potential (or registered) voters assigned to treatment. For interventions that would occur in the absence of an experiment,  $\mathcal{S}_d$  represents the proportion of potential voters that would (resp. would not) have been exposed to the intervention due to experimental assignment of treatment.

### 3.2 Researcher Assumptions about Interference between Voters

To construct bounds on interference between individuals and clusters, researchers must make some assumptions about the set of voters impacted by an intervention. First, consider the stable unit treatment value assumption (SUTVA), which is typically invoked to justify identification of causal estimands in experimental research. In the setup from the previous section, this means that a

voter’s potential outcomes are independent of the assignment of any other voter outside her cluster, where the cluster represents the unit of assignment as defined above. Denoting a binary treatment,  $Z \in \{0, 1\}$ , SUTVA for electoral outcome  $Y_j(z_{jc})$  is written in Assumption 1.

**Assumption 1.** *SUTVA:*  $Y_{jc}(z_{jc}) = Y_{jc}(z_{jc}, \mathbf{z}_{j, \neg c})$

I add a second *within-cluster* non-interference assumption to the baseline assumption. Note that, in contrast to SUTVA, this assumption is not necessary for identification of the average treatment effect (ATE) in cluster-randomized experiments. This assumption holds that, in the case that treatment is assigned to clusters of more than one voter ( $n_c > 1$ ), a voter’s potential outcomes are independent of the assignment of any other voter inside her cluster, where the cluster represents the unit of assignment to treatment.<sup>4</sup> I express this assumption formally in Assumption 2. In other words, Assumption 2 holds that an intervention could only influence the voting behavior of voters directly allocated to receive the intervention. Analysis of within-cluster “spillover” effects in experiments suggest that this assumption is not always plausible in electoral settings (i.e., Ichino and Schündeln, 2012; Sinclair, McConnell, and Green, 2012; Giné and Mansuri, 2018), so I examine the implications of relaxing this assumption in Section 5.

**Assumption 2.** *No within-cluster interference:*  $Y_{jc}(z_{jc}) = Y_{jc}(z_{jc}, \mathbf{z}_{\neg j, c})$

### 3.3 Voters’ Response to the Treatment

Because the question at hand relates to whether an experimental intervention can change aggregate election outcomes, I focus on voting outcomes. To accommodate the range of interventions in the literature, I assume the potential outcomes framework as a tractable and agnostic model of voting behavior for bounding outcomes. Specifically, given a treatment  $z \in Z$ , I assume that a vote choice potential outcome  $A_{jc}(z) \in \{0, 1\}$  is defined for all  $j, z$ , where 1 corresponds to a vote for the marginal (*ex-ante*) winning candidate and 0 represents any other choice (another candidate, abstention, an invalid ballot, etc.).

---

<sup>4</sup>This assumption holds trivially in individually-randomized experiments when  $|n_c| = 1$  or when all registered voters in a cluster are treated.

I bound the plausible treatment effects on vote choice for the marginal “winner” among those whose assignment to treatment is changed by the use of an experiment, i.e. any  $j \in \{S_{10} \cup S_{01}\}$ . Given the binary vote choice outcome, one can bound the possible (unobservable) individual treatment effects, among subjects whose treatment status is changed through the use of an experiment as:  $ITE_{jc} \in \{-a_{jc}(0), 1 - a_{jc}(0)\}$ . If a voter would vote for the winner when untreated ( $a_{jc}(0) = 1$ ), she could be induced to vote for a different candidate  $-a_{jc}(0) = -1$  or continue to support the winner  $1 - a_{jc}(0) = 0$  if treated. Conversely, if a voter would not vote for the winner if untreated ( $a_{jc} = 0$ ), her vote (for any non-winning candidate) could remain unchanged  $-a_{jc}(0) = 0$  or she could be induced to vote for the winner  $1 - a_{jc}(0) = 1$  if treated. Note that these effects serve as the basis for construction of extreme value bounds (EVB) (Manski, 2003).

EVBs can be very wide. Researchers may instead be tempted to use estimates from related studies or less conservative bounding approaches. These approaches may be very misleading. Existing estimates are generally some form of average causal effect (i.e., an *ATE*). The ethical concern is not whether an experimental intervention changes outcomes on average. If *ATE* estimates that are small in magnitude mask heterogeneity, bounds based upon existing estimates will be non-conservative. Moreover, the outcome measures on vote share for a specific party or incumbent party in the existing literature likely do not uniformly correspond to the relevant *ex-ante* marginal winning candidates, which means that they do not measure the effect that directly corresponds to considerations of aggregate electoral impact. Relative to other bounding approaches, I adopt EVB to avoid incorrect assumptions about the plausible effects of an intervention (i.e., monotonicity). Indeed, Bayesian models of voter updating invoked in information experiments predict non-monotonicity in treatment effects as a function of the location of the signal relative to the prior (i.e. “good” vs. “bad” news).

Voting outcomes are observed at the level at which treatment is assigned, indicated by  $c$ . I assume that in treatment clusters where  $n_c > 0$ , registered voters are randomly sampled to receive the intervention. The expectation of untreated potential outcome  $E[a_c(0)]$  plays an important role in the construction of bounds on aggregate electoral impact. Random sampling ensures that  $E[a_c(0)]$

is equivalent at varying levels of experimental saturation in a treated cluster. This assumption can be relaxed when it is inappropriate, but the bound on aggregate electoral impact may increase.

## 4 Bounding Effects on Electoral Behavior

### 4.1 Bounding Electoral Impact

Given the design elements characterized by the (experimental) assignment of treatment, researcher assumptions about interference, and the model of voter response to treatment, I proceed to construct an *ex-ante* bound on the largest share of votes that could be changed by an experimental intervention. I term this term, the *maximum aggregate electoral impact* in a district, the  $MAEI_d$ . Under Assumptions 1 and 2, this quantity is defined, by electoral district, as:

**Definition 1.** *Maximal Aggregate Electoral Impact: The ex-ante maximal aggregate electoral impact (MAEI) in district  $d$  is given by:*

$$MAEI_d = \max \left\{ \frac{\sum_{c \in d} [E[a_c(0)] | S_{10}^{cd} \cup S_{01}^{cd}|]}{n_d}, \frac{\sum_{c \in d} [(1 - E[a_c(0)]) | S_{10}^{cd} \cup S_{01}^{cd}|]}{n_d} \right\} \quad (2)$$

Consider the properties of  $MAEI_d$  with respect to untreated levels of support for the winning candidate. Note that  $E[a_c(0)] \in [0, 1]$  for all  $c \in d$ . This has two implications. First, because  $E[a_c(0)]$  is unknown *ex-ante*, a conservative bound can always be achieved by substituting  $E[a_c(0)] = 1$  (equivalently 0). These conservative bounds are useful when the intervention is assigned to a non-random sample of registered voters within a cluster. Second, holding constant the experimental design, the  $MAEI_d$  is minimized where  $E[a_c(0)] = \frac{1}{2}$  for all clusters in a district, with non-empty  $S_{10}^{cd}$  or  $S_{01}^{cd}$ . Thus, going from the least conservative prediction of  $E[a_c(0)] = \frac{1}{2}$  for all  $c$  to the most conservative assumption of  $E[a_c(0)] = 1$  for all  $c$ , the magnitude of  $MAEI_d$  doubles.

Inspection of Definition 1 yields several observations. Most obviously, an identical experiment has less possibility of moving aggregate vote share or turnout in a large district relative to a small district. In other words, the bounds we can place on the electoral impact of the same experimental

design are much narrower for a presidential election than for a local school board election. Note that researchers' desire to work in low-information contexts has directed research focus to legislative or local elections. This result suggests that this decision carries greater risks of changing electoral outcomes, all else equal.

Second, Definition 1 implies that a higher saturation of the experimentally-manipulated treatment increases the potential for effects on vote share, holding constant  $E[a_c(0)]$ . This suggests a trade-off between statistical power and the degree to which an experiment could alter aggregate electoral outcomes. Increasing the saturation of treatment introduces the possibility of changing more votes. For example, changing from an individually-randomized to a cluster-randomized experiment requires many clusters for adequate power to detect treatment effects. When researchers treat large proportions of voters in cluster-randomized experiments, the saturation of treatment increases substantially. Thus, when individual-level outcome data is not available, researchers compensate by cluster-randomizing treatment, often increasing the  $MAEI_d$  substantially.

## 4.2 Assessing the Consequences of Electoral Interventions

The implications of  $E[a_c(0)]$  on  $MAEI_d$  prompt a discussion of the ability of electoral experiments to change electoral outcomes, that is, who wins. While analyses of electoral experiments typically focus on vote share, not probability of victory (or seats won in a proportional representation system), the lever through which elections have consequences is who wins office.

The mapping of votes to an office or (discrete) seats implies the existence of at least one threshold, which, if crossed, yields a different realization of office holding. For example, in a two candidate race without abstention, there exists a threshold at 50 percent. It is useful to denote the “*ex-ante* margin of victory,”  $\psi_d$ , as minimum change in vote share, as a proportion of registered voters, at which a different officeholder would be elected in district  $d$ . In a plurality election for a single seat, the margin of victory terminology is obviously familiar. In a proportional representation (PR) system, there are various interpretations of  $\psi_d$ . Perhaps the most natural interpretation is the smallest change in any party's vote share that would change the distribution of seats.

If  $\psi_d > 2MAEI_d$ , then an experiment could not change the ultimate electoral outcome. In



contrast, if  $\psi_d < 2MAEI_d$ , the experiment *could* affect the ultimate electoral outcome. Appendix A1 shows formally the derivation of this threshold for an  $n$ -candidate race. The intuition behind the result is straightforward:  $n_d\psi_d$  gives the difference in the number of votes between the marginal winning and losing candidates. The minimum number of votes that could change the outcome is  $\frac{n_d\psi_d}{2}$  (assuming a fair tie-breaking rule), if all changed votes were transferred from the marginal winner to the marginal loser. Hence, the relevant threshold is  $2MAEI_d$ , not simply  $MAEI_d$ .

Unlike the other parameters of the design,  $E[a_c(0)]$  and  $\psi_d$  are not knowable in advance of an election, when researchers plan and implement an experiment. Imputing the maximum possible value of  $E[a_c(0)] = 1$  allows for construction of the most conservative (widest) bounds on the electoral impact of an experiment under present assumptions, maximizing  $MAEI_d$  while fixing other aspects of the design. However, imputing the minimum value of  $\psi_d = 0$ , the most “conservative” estimate, implies that  $2MAEI_d > \psi_d$  and *any* experiment could change the electoral outcome. Yet, we know empirically that not all elections are close and, in some settings, election outcomes can be predicted with high accuracy. For this reason, bringing pretreatment data to predict these parameters allows researchers to more accurately quantify risk and make design decisions.

To this end, researchers can use available data to predict the parameters  $\psi_d$  and, where relevant,  $E[a_c(0)]$ . Given different election prediction technologies and available information, I remain agnostic as to a general prediction algorithm. Regardless of the method, we are interested in the predictive distribution of  $\psi_d$ ,  $\hat{f}(\psi_d) \sim f(\psi_d|\hat{\theta})$ , where  $\hat{\theta}$  are estimates of the parameters of the predictive model.

### 4.3 Decision Rule: Which (if Any) Experimental Design Should be Implemented?

Ultimately, our assessment of whether an experimental design is *ex-ante* consistent with the ethical standard of not changing aggregate electoral outcomes requires a decision-making rule. I propose the construction of a threshold based on the predictive distribution of  $\psi_d$ . In particular, I suggest that researchers calculate a threshold  $\underline{\psi}_d$ , that satisfies  $\hat{F}^{-1}(0.05) = \underline{\psi}_d$ , where  $\hat{F}^{-1}(\cdot)$  indicates the quantile function of the predictive distribution of  $\psi_d$ . This means that 5% of hypothetical realizations of the election are predicted to be closer than  $\underline{\psi}_d$ . The decision rule then compares

$MAEI_d$  to  $\underline{\psi}_d$ , proceeding with the experimental design only if  $2MAEI_d < \underline{\psi}_d$ .

This decision rule rules out intervention in close elections entirely. It permits experiments with a relatively high experimental saturation of treatment only in predictable “landslide” races. Moreover, basing a decision rule on predictive distribution of  $\psi_d$  as opposed to the point prediction,  $\widehat{\psi}_d$  penalizes uncertainty over the possible distribution of electoral outcomes. Globally, the amount of resources and effort expended on predicting different elections is vastly unequal. As a result, we are able to make relatively more precise predictions in some races than others.

Substantively, this decision rule corresponds to a determination not to change any individual voter’s ability to be pivotal. Variation in pivotality represents one source of political inequality across any electorate. Adherence to the proposed decision rule simply circumscribes researchers’ ability to change (in either direction) the pivotality of a subject or non-subject in an electoral district, limiting researchers’ ability to change the distribution of political power. This connection between pivotality and political equality speaks to concerns enumerated by Beerbohm, Davis, and Kern (2020) about electoral experiments generating various forms of political inequality.

## 5 Allowing for Spillovers/Interference

Due to the use of extreme value bounds, decisions based on the  $MAEI_d$  are conservative under the assumptions on interference posited Section 3.2. By conservative, I mean that they will induce a researcher to err on the side of not conducting the experiment. Yet, when these assumptions do not hold, the same analysis might justify a non-conservative decision. For this reason, I examine the implications of relaxing these assumptions.

### 5.1 Within-Cluster Interference

One limitation of the previous analysis, is that an intervention might only change the votes of those that are directly exposed within a cluster (Assumption 2). In this instance, clusters consist of multiple voters ( $n_c > 1$ ) but not all voters in a treated cluster are treated or untreated due to the experiment. Yet, some “always assigned” (where present) or “never assigned” voters in assigned clusters may change their voting behavior in response to the treatment administered to other vot-

ers in their cluster. In electoral context, these spillovers may occur within households (Sinclair, McConnell, and Green, 2012), intra-village geographic clusters (Giné and Mansuri, 2018), or constituencies (Ichino and Schündeln, 2012). In these cases, the maximum aggregate electoral impact with within-cluster interference,  $MAEI_d^w$  can be rewritten as:

$$MAEI_d^w = \max \left\{ \frac{\sum_{c \in d} [E[a_c(0)]n_c I[|S_{10}^{cd} \cup S_{01}^{cd}| > 0]]}{n_d}, \frac{\sum_{c \in d} [(1 - E[a_c(0)])n_c I[|S_{10}^{cd} \cup S_{01}^{cd}| > 0]]}{n_d} \right\} \quad (3)$$

where  $I[\cdot]$  represents an indicator function. Note that the bound  $MAEI_d^w$  maintains SUTVA (Assumption 1).

Two elements change from  $MAEI_d$  to  $MAEI_d^w$ . First, the number of voters whose potential outcomes may be affected by the experimental intervention increases to include all voters in a cluster in which any voter's assignment status is changed by an experiment. This follows from the fact that  $|S_{10}^{cd} \cup S_{01}^{cd}| \leq n_c$ . Second, the expectation of untreated turnout,  $E[a_c(0)]$  is now evaluated over all registered voters in a cluster (not just subjects). In the context of randomized saturation designs,  $E[a_c(0)]$  does not change because the cluster population is randomly sampled. Random sampling within a cluster is sufficient to ensure that  $MAEI_d^w \geq MAEI_d$ . In other words, within-cluster interference increases the size of the possible electoral impact of an intervention. This analysis implies that if the only form of interference is within-cluster, we can construct a conservative bound on the aggregate impact of an experiment without further assumptions.

## 5.2 Between-Cluster Interference

I now to proceed to relax SUTVA, Assumption 1. Note that SUTVA is typically assumed to justify identification in electoral experiments. In order to account for between-cluster interference, a violation of SUTVA, I introduce a vector of parameters  $\pi_c \in [0, 1]$ , indexed by  $c$ , to measure researchers' *ex-ante* beliefs about the proportion of voters that could respond to treatment (or some manifestation thereof) in clusters where allocation of the intervention is not changed by the experiment. In experiments in which the intervention would not occur absent the experiment, this

term refers to the set of registered voters in control clusters.

$$MAEI_d^{bw} = \max \left\{ \frac{\sum_{c \in d} [E[a_c(0)]n_c I[|S_{10}^{cd} \cup S_{01}^{cd}| > 0] + E[a_c(0)]n_c \pi_c I[|S_{10}^{cd} \cup S_{01}^{cd}| = 0]]}{n_d}, \right. \\ \left. \frac{\sum_{c \in d} [(1 - E[a_c(0)])n_c I[|S_{10}^{cd} \cup S_{01}^{cd}| > 0] + (1 - E[a_c(0)])n_c \pi_c I[|S_{10}^{cd} \cup S_{01}^{cd}| = 0]]}{n_d} \right\} \quad (4)$$

The new term in the numerator of both expressions in Equation 4 reflects the possible changes in turnout in clusters where no subjects' assignment to the intervention is changed due to the experiment. Intuitively, because  $\pi_c \geq 0$ , it must be the case that the aggregate electoral impact of experiments that experience between- and within-cluster interference is greater than those with only within-cluster interference,  $MAEI_d^{bw} \geq MAEI_d^w$ .

Now, consider the implications of conservatively setting  $\pi_c = 1$  for all  $c$ , akin to an assumption that an experiment could affect the potential outcomes of all registered voters in a district. In this case, Equation 4 simplifies to:

$$MAEI_d^{bw} = \max \left\{ \frac{\sum_{c \in d} E[a_c(0)]n_c}{n_d}, \frac{\sum_{c \in d} (1 - E[a_c(0)])n_c}{n_d} \right\} \text{ if } \pi_c = 1 \forall c \quad (5)$$

However, it must always be case that the *ex-ante* margin of victory,  $\psi_d \leq \frac{1}{n_d} \sum_{c \in d} E[a_c(0)]n_c$ , as this represents the case in which the winning candidate wins every vote. It therefore must be the case that if  $\pi_c = 1 \forall c$ ,  $\psi_d \leq 2MAEI_d^{bw}$ . In other words, without circumscribing  $\pi_c$  in some way, we would never satisfy the decision rule proposed in this article in a contested election. As such, a researcher should never run an electoral experiment if she anticipates between-cluster spillover effects that could reach all voters, even absent problems of identification and inference.

### 5.3 General Equilibrium Effects

The discussion of interference has been agnostic as to the mechanism for between or within-cluster interference. Because of the need to bound  $\pi_c$ , it is useful to consider why more voters may be exposed to some manifestation of the experimental intervention. The causal estimands identified

by electoral experiments are generally motivated (explicitly or non-explicitly) as tests of “partial equilibrium” comparative statics in which voters respond to a treatment in isolation. However, other actors – typically candidates, campaigns, or other voters – may also respond to an intervention in attempts to win elections. Such actions change: (1) the treatment bundle received by voters; and (2) the set of voters that receive any part of that bundle. For the researcher designing an experiment, the validity of the present bounding exercise depends on foresight into the set of actors that could respond to treatment and the actions they might take.

Examination of the literature suggests that reactions by other actors can increase or decrease the share of voters exposed to the intervention through the experiment. For example, in an accountability experiment in India, the detention of field staff by affiliates of a candidate and eventually local police curtailed the intervention after less than 10% of the intervention period (Sircar and Chauchard, 2019). In this sense, “general equilibrium” effects ended the intervention, leading to many fewer treated voters than the researchers planned. On the opposite extreme, a postcard intervention insinuating candidate partisanship in a non-partisan Montana judicial election drew the ire of state officials and the attention of national press, plausibly exposing far more than the 14.8% of Montanan registered voters assigned to the intervention to some manifestation thereof.<sup>5</sup>

To the extent that scholars have measured campaign response to voter-level experimental treatments, works like Arias et al. (2019) suggest that incumbents and challengers did choose to amplify or mitigate informational disclosures in an accountability experiment. Importantly, such actions are not precisely targeted to treated voters, suggesting that such responses exposed more voters to some manifestation of the intervention than did the researchers. This suggests some  $\pi_c > 0$ , though the plausible range of effects consistent with these measurements is small. Note that if outside actors accurately target general equilibrium responses inside treatment clusters, the bound in Equation 3 is conservative. If, however, such targeting reaches untreated voters outside the cluster, the bound widens. Most challengingly, such a determination must be made before the intervention is fielded.

---

<sup>5</sup>This calculation is based on report of 100,000 fliers in Michelson (2016).

## 6 Applications

To this point, the framework developed is silent with regard to the *content* of experimental treatments. Following the classification of interventions in Figure 1, I consider whether there are specific interventions in which the decision rule is unnecessarily conservative.

It is useful to consider which registered voters’ behaviors might be impacted by an intervention. Indeed, of existing interventions intend to change the behavior of a subset of the electorate. For example, mobilization (get out the vote) interventions theoretically aim to change the behavior of registered voters who would not otherwise turn out to vote. Persuasion experiments aim to change the votes of individuals that intend to vote for a different alternative. Noting these distinct aims, recall that the decision rule,  $2MAEI_d < \underline{\psi}_d$ , allows for the fact that voters may switch their votes between the *ex-ante* marginal winner and loser. If a mobilization intervention only mobilized non-voters to vote (or demobilized voters from voting), a less demanding decision rule of  $MAEI_d < \underline{\psi}_d$  would be appropriate for mobilization interventions. Yet, this less stringent decision rule makes the (potentially strong) assumption that mobilization interventions do not affect the vote choice of subjects who would vote regardless. We have little evidence for or against this assumption. Due to the secret ballot, we cannot measure persuasive effects at the individual level using voter file data, the most common outcome data source in mobilization experiments (Green and Gerber, 2015). In contrast, voter registration experiments (e.g., Shineman, 2020) comprise an emerging intervention class that may productively invoke the less stringent  $MAEI_d < \underline{\psi}_d$  rule. Specifically, if these interventions can be directed exclusively to unregistered individuals, we mechanically eliminate the possibility of “persuasive effects” among subjects.

### 6.1 Implementing the framework

I examine the application of this framework to existing experiments in two ways. In Appendix A4, I use replication datasets, administrative, and archival data to apply the framework to four published experiments that comprise mobilization, information, and persuasion interventions: Boas, Hidalgo, and Melo (2019), Gerber and Green (2000), Bond et al. (2012), and López-Moctezuma

et al. (2021). Further, in Appendix A5, I present back-of-the-envelope calculation of the  $MAEI_d$  on 14 studies of information and accountability that are classified by Enríquez et al. (2019). This analysis shows that these research designs vary substantially  $MAEI_d$  and justify the considerations I forward. Moreover, I show that eight of the 14 studies do not report information on how the experimental units relate (quantitatively) to the electorate as a whole (at least outside of replication materials). This occurs either because: units (voters or clusters) were not randomly sampled from the district (4 studies) or because there is insufficient information about constituency size,  $n_d$  (4 studies). This non-random sampling is generally well-justified from a design perspective and constituency size is not necessary for the estimation of causal effects. However, to map these research designs into the present framework or onto electoral outcomes generally, we would need information relating the experimental sample to the district(s) in which they are conducted. The takeaway from this survey of 14 studies is simply that considerations of aggregate electoral impact require analyses that are not standard practice.

Simulating the design of experiments under the decision rule allows for an application of the full framework. The simulation below uses electoral data from the US state of Colorado. It relies upon real voter registration data, precinct-to-district mappings, and election predictions. Because US elections are administered at the state level, the simulations are greatly simplified by focusing on a single state in one election: the 2018 midterms. All races in 2018 were at the state level or below. In the simulations, I assume that an experimental intervention would not occur absent the researcher. There are many forecasts available for the 2018 US House elections. I use the forecast by Morris (2018). I do not know of forecasts for Colorado State House seats. Therefore, I predict outcomes from (limited) available data, namely partisan voter registration data and lagged voting outcomes. I fit a basic predictive model on electoral data from 2012-2016 (three elections) and then predict outcomes for 2018 (see A6.3 for details).

Examining only the predictive intervals, Figure 2 depicts the 90% predictive intervals for Colorado’s 65 State House and 7 US House seats in 2018. The 90% predictive intervals provide a useful visualization because when they bound 0 (gray intervals in the Figure), no experiment can

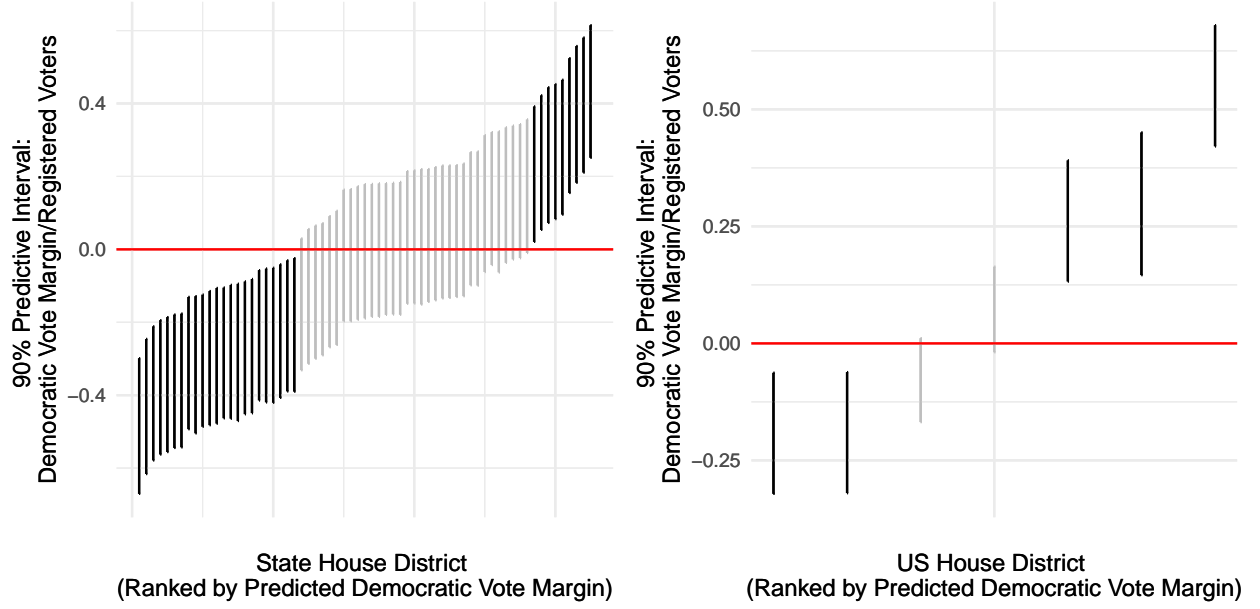


Figure 2: Predictive intervals for 65 State House seats and 7 US House seats. Gray intervals represent grounds for declining to conduct an experiment in a district under the decision rule proposed here.

pass the decision rule proposed in this paper. In sum, 33/65 State House races and 2/7 US House races bound 0.

I consider two research designs, each invoking SUTVA and, by design, satisfying Assumption 2.<sup>6</sup> I first consider experiments that assign individual voters (not clusters) to treatment. I show calculations based on three types sampling of individuals into the experimental sample that vary the calculation of  $E[a_c(0)]$  and thus  $MAEI_d$ . A best case scenario sets  $E[a_c(0)] = \frac{1}{2}$  and represents the case in which participants were pre-screened to evenly fall on both sides of the ideological spectrum. A worst case scenario sets  $E[a_c(0)] = 0$  (resp. 1) and could represent the case in which all experimental subjects would vote in the same way absent treatment, which approximates a sample composed of only strong partisans. The intermediate case represented by “random sampling” predicts  $E[a_c(0)]$  from 2016 district vote totals.

Figure 3 depicts the theoretical maximum number of individuals that could be assigned to *treatment* in State House and US House elections, by district and race. The shading represents the three

<sup>6</sup>I assume all voters in cluster-randomized precincts are assigned to treatment if they belong to a treated cluster.



sampling assumptions described above. Several features are worth note. First, the experimental allocation of treatment can only pass the decision rule in sufficiently extreme (thus predictable) electorates. Ranking districts from the most Republican to most Democratic (in terms of predicted vote margin) on the  $x$ -axis, the maximum number of individuals assigned to treatment is 0 in competitive races. The more lopsided the race, the more subjects can be assigned to treatment under the decision rule. Second, the type of experimental sample conditions the permissible treatment group size. However, going from worst to best case can double the number of subjects, as implied by Equation 2. Third, comparing the top and bottom plots in the left column, in larger districts, the maximum number of registered voters that could be assigned to treatment grows proportionately to district size (see Table A8 for summary statistics). Finally, when describing the maximum number of treated subjects as a proportion of the electorate, only sparse treatments are permissible under the decision rule. Nevertheless, it implies that one could allocate an individually-randomized treatment in a way such to power an experiment within the ethical constraints proposed by this article. Figure A11 reports the results of an analogous cluster-randomized treatment at the *precinct* level revealing that the number of “treatable” precincts is quite small, particularly in the case of State House races.

## **7 Implications for Research Design and Learning**

The parameters used to characterize elections reflect features of both electoral systems, context, and data availability. I argue that best practices for electoral experiments are more likely to be tenable in some institutional contexts than others, as enumerated in Table 4.

These institutional features, and a number of contextual features, may circumscribe the use of electoral experiments. The framework developed here suggests a need for selection on intervention content and features of elections, yielding five recommendations for design of these experiments:

1. Select treatments to improve the plausibility of assumptions of restricted interference.
2. Experiment in FPTP races.
3. Implement interventions in larger electoral districts.

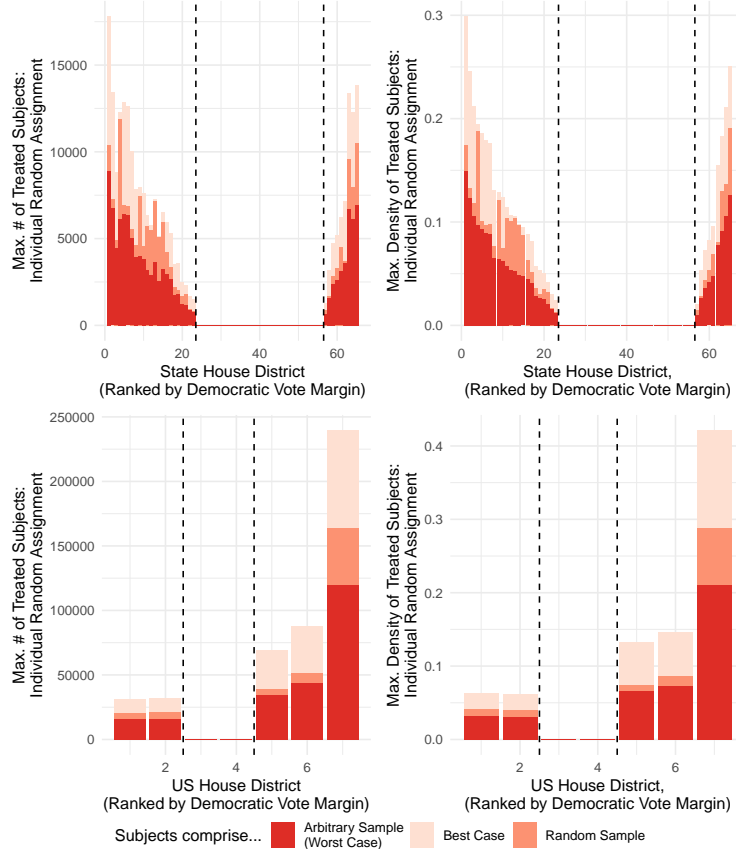


Figure 3: Maximum number of individuals (left) or proportion of registered voters (right) that can be assigned to treatment under decision rule.

Parameter	Feature of elections	Implications
$\psi_d$ (margin of victory)	<p><b>Electoral systems</b> change the interpretation of margin of victory. In FPTP races, it is readily interpretable as the difference in vote share of the top two candidates. In PR races, <math>\psi_d</math> could be interpreted in terms of the last seat allocated <i>or</i> the allocation of seats within a list.</p> <p><b>District magnitude</b> constrains the possible range of <math>\psi_d</math>. In single-round systems, <math>\max\{\psi_d\} = 1/\text{District magnitude}</math>.</p>	<p>Limits to our ability to interpret margin of victory may limit the ability to predict this quantity precisely, which limits the possibility that the decision rule is satisfied.</p> <p>Increases in proportionality under proportional representation limit the possibility of “land-slide” elections where moderate-density treatments would be unable to move outcomes.</p>
$n_d$ (number of registered voters)	<p><b>Concurrent elections</b> may imply that an intervention on a set of voters may represent a much larger proportion of the electorate in one race than in a concurrent race.</p>	<p>Concerns can be minimized if the experimental manipulation happens in the “smallest” race: the race with the smallest <math>n_d</math>. Concurrent elections can lead to large differences in assessments of the risk of electoral experiments.</p>

Table 4: Features of electoral systems and mapping to the framework. Note that we may district magnitude and  $n_d$  to be negatively correlated, with countervailing implications for calculation of  $MAEI_d$ . However, that at least in some countries like Brazil, local elections include both FPTP races for executives and PR races for local councils with the same electorate (holding fixed  $n_d$ ).

4. Avoid implementing experimental interventions in close or unpredictable races.
5. When intervening without a partner, reduce the number of subjects assigned to treatment in each district,  $d$ . When intervening with a partner, reduce the number of voters assigned to treatment or control because of the random assignment in each district,  $d$ .

These recommendations complement and extend guidance from Desposato (2016) that advocates reductions in sample size (#5) and consideration of pre-race polling where available (#4). Desposato (2016) further advocates a power calculation to aid in minimizing the sample size in service of ethical concerns. The method in this article suggests that the level of treatment saturation implied by a power analysis can be too risky to implement. Moreover, I provide additional design levers—not simply sample size—that researchers can use to mitigate the possibility of changing aggregate election outcomes.

These design strategies posit trade-offs in terms of learning from electoral experiments. I focus on implications for generalization (external validity) and statistical power. Strategy #1 circumscribes the set of treatments that researchers develop and administer experimentally. In particular, this paper suggests that treatments that vary saturation of treatment assignment to study social dynamics or network effects of voting behavior are unlikely to pass the decision rule. Certainly some social dynamics like coordination might be examined experimentally within small subsets of the electorate (i.e., within the household), but the current emphasis on these dynamics within large subsets of the electorate yields designs that are less likely to pass the decision rule forwarded here. More attention to when these types of interference are most likely can inform the choice of interventions examined in electoral experiments.

Strategies #2 and #3 constrain the types of races in which intervention consistent with the ethical guidelines in this article is feasible. These strategies rule out electoral experimentation in some countries or offices as a function of the electoral system or institutions. If voters, candidates, and parties behave differently under different electoral systems, it may be hard to generate evidence relevant to different types of elections. To the extent that framework focuses on the ratio of treated voters to registered voters, the guidance to experiment in larger districts is similar to the guidance

to reduce the number of voters assigned to treatment. However, it also speaks to the motivation of the interventions attempted. Many mobilization interventions focus on lower-turnout elections (i.e., primaries, state, or midterm elections) and many information interventions focus on low-information elections (often local races). These approaches contrast with the guidance of Strategy #3. Note that both approaches—the current emphasis on low-turnout or low-information elections and the recommendations to focus on races in larger districts—speak to potential limitations of generalization from electoral experiments. While critiques of the lack of external validity of experiments are widespread, the idea that ethical considerations may lead to a less sample that is less “representative” is new to my knowledge.

Strategy #4—avoiding close or unpredictable races—posits concerns for both generalization and statistical power. This strategy excludes some polities with high volatility or minimal investment in election prediction. A discussion of limited external validity in the context of experiments implies a concern about treatment effect heterogeneity. Indeed, in electoral contexts, we may expect voters (or politicians) to act differently in places where a voter is more or less likely to be pivotal. If treatment effects vary in the characteristics used to target an experimental intervention, there exists a trade-off between these recommendations and the generalizability of insights about behavior. A focus on landslide races may also impact statistical power, though the direction is ambiguous. If there are fewer persuadable (“swing”) voters in landslide districts than in marginal districts, a lack of persuadable voters may imply a lower ceiling for treatment effects, limiting power. However, power also depends on the distribution of the outcome variable. For a binary (voter-level) outcome, power to detect a (fixed) effect size is higher in very imbalanced electorates. The net implications for power are therefore ambiguous.

Finally, Strategy #5 points to a familiar trade-off between statistical power and concerns about impacting aggregate electoral outcomes. I show that this trade-off is particularly salient in experiments seeking to analyze aggregate electoral outcomes at the cluster (i.e., polling station or precinct) level. At the same time, the framework provides novel guidance for the allocation of treatment across electoral districts. Specifically, it suggests that some power concerns may be reduced

through higher levels of treatment saturation in districts where elections are highly predictable and not close and lower levels of treatment saturation in districts where elections are predicted to be somewhat more competitive. As such, the framework provides new ways to improve statistical power given this known trade-off.

Does the circumscription of electoral experiments to certain electoral contexts and treatments undermine the utility of electoral experiments as a tool? Here, an analogy to electoral regression discontinuity designs (RDDs) proves instructive (Lee, 2008). Electoral RDDs estimate some form of local average treatment effect at the threshold where elections are decided. The method is disproportionately used in low-level (i.e. municipal) FPTP contests, in search of statistical power and questions about how to conceptualize the running variable in proportional representation contests (but see Folke, 2014). If these limitations on the application of electoral experiments are to be seen as damning to electoral experiments but not electoral RDDs, there seemingly exists a question of whether the study of landslide races are less interesting—or of less political importance—than close contests. Theoretically, there are reasons why close contests may reveal distinct strategic dynamics that are not evident in predictable landslides, but this claim seems non-obvious. As such, this article simply advocates for a more careful application of electoral experiments with broader recognition of their limitations, not a wholesale abandonment of the tool.

## **8 Discussion: Revisiting the Ethical Objective**

Readers of this article may object to the premise that electoral experiments should be designed to avoid changing aggregate electoral outcomes, though for different reasons. Some readers may argue that social scientists should not be intervening in real elections at all, holding that the ethical objective here is too permissive. Other readers may argue that social scientists are sometimes justified in changing election outcomes, holding that the ethical objective here—and in existing literature—is unnecessarily constraining. I discuss both sets of arguments.

## 8.1 Objections part I: The ethical objective is too permissive

Avoiding changing aggregate electoral outcomes is clearly not the only ethical consideration that should arise when designing an electoral experiment. Researchers should address standard ethical considerations around intervention and measurement in addition to the considerations that I develop in this paper. However, two such considerations—consent and self-determination—might be argued to supersede considerations of aggregate electoral impact and call for stronger restrictions (or bans) on the use of electoral experiments.

**Lack of consent:** Like most field experiments, electoral experiments are generally conducted without the consent of subjects or non-subjects who may be affected by the intervention.<sup>7</sup> Teele (2013), Humphreys (2015), and McDermott and Hatemi (2020) note that the lack of consent in field experiments departs from standard requirements of informed consent in medical studies. Desposato (2018) shows that informed consent increases the proportion of American survey respondents and political scientists that find hypothetical experiments ethically acceptable. Objections to field experimentation on the basis of lack of informed consent extend far beyond electoral experiments. Humphreys (2015) and Teele (2019) chart a productive path forward concerning issues of consent. By conceptualizing consent more broadly, both authors provide new suggestions about how experimentalists might seek consent in new ways or provide additional protection to subjects and non-subjects in its absence. Further development of and debate about these alternatives is important to addressing issues of consent as a broader objection to field experimentation.

**Electoral experiments violate self-determination:** This paper makes a consequentialist argument: experiments on elections can generate social harm by changing who wins office. Baele (2013: p. 28) asserts that the primary deontological issue with electoral experiments is that they “influenc[e] political situations in other countries ... as it constitutes a breach in sovereignty if all the stakeholders do not agree in the process.” Whitfield (2019: p. 7) argues more broadly for political research ethics that respect “self-determination of communities.” As such, even if an ex-

---

<sup>7</sup>While some researchers obtain the consent of (select) candidates or political parties, I am not aware of field-experimental electoral interventions that seek the consent of voters (subjects or non-subjects).

periment were to pass the decision rule I advance, violations of an community's (electorate's) right self-determination may justify broader restrictions on electoral experimentation. Two caveats are important. First, to my knowledge, this argument has not yet been developed in depth with specific reference to electoral experiments. Future work is needed to justify this highly restrictive stance on electoral experimentation. Second, it is unclear how these arguments apply to a researcher that implements an intervention in their own electoral district. Relatedly, Beerbohm, Davis, and Kern (2020) provides useful guidance on the moral status of treatments carried out by a researcher (presumably in their home district).

## **8.2 Objections part II: The ethical objective is too strict**

In sharp contrast, some readers may contend that the ethical guidance to not change election outcomes is too restrictive. Indeed, I have shown that there exist institutional and political contexts in which no experimental design is likely to pass the proposed decision rule. These objections largely center two arguments. First, there may be benefits—either in the form of welfare or knowledge—that stem from electoral experiments. Second, the political science literature suggests that election outcomes may have many causes. Why, then, should we care so much about one (possible) cause: electoral experiments?<sup>8</sup>

**The benefits of electoral experiments:** Electoral interventions generate learning benefits that may also generate welfare benefits for subjects and their communities (Davis and Michelitch, 2021). Importantly, many scholars are motivated by problems that plague elections including electoral fraud, clientelism, and underrepresentation of some groups on ballots or at the ballot box. These interventions—or knowledge gained from these interventions—may provide immediate benefits to subjects and their communities or, in the longer term, benefits to others via learning.

The principle of beneficence holds that researchers should “maximize possible benefits and minimize possible harms” (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research., 1979: p. 5). Changing an electoral outcome will harm some in-

---

<sup>8</sup>Note that electoral experiments allow us to measure effects of causes, not causes of effects. Yet, we can still theorize about possible causes of these election outcomes.

dividuals while benefiting others (Gubler and Selway, 2016; Zimmerman, 2016). For example, a candidate that loses *because of the intervention* and her supporters will generally be harmed by an intervention. By contrast, a candidate that wins *because of the intervention* and her supporters will generally benefit from the intervention. Researchers are often unable to anticipate or weigh the extent of the harms and benefits across an electorate (Carlson, 2020; Baele, 2013). When researchers lack foresight into the consequences of the intervention, they can minimize the potential for harm—while still gaining knowledge benefits—by designing the experiment to minimize aggregate electoral impact following the guidance in this article.

But what if a researcher believes *ex-ante* that their treatment will improve welfare by changing an election outcome? Here, consideration of beneficence demands asking whether random assignment of treatment—a necessity for experimentation—is consistent with maximizing the potential benefits of the intervention. By withholding treatment in order to create a control group, a researcher necessarily limits their ability to produce the welfare-improving outcome. Even when an intervention is scarce, targeting treatment non-randomly (i.e., to likely “swing” voters) rather than to a random cross-section of voters is likely to be more efficient in achieving the welfare-enhancing outcome. But this non-random assignment eliminates the experiment and comprises learning.

This discussion suggests that the simultaneous pursuit of knowledge and welfare gains is possible under four conditions: (1) there exists a welfare-improving outcome that is known *ex-ante*; (2) the intervention is known to increase the likelihood of that outcome; (3) treatment is scarce and so not all relevant actors can be treated; and (4) researchers lack sufficient information about subjects to target the treatment more efficiently than random assignment. When these conditions obtain simultaneously, researchers may be justified in implementing a design liable to change aggregate outcomes. However, these conditions are extraordinarily restrictive and should be justified when this argument is invoked. The starkness of these conditions suggests that the objective proposed here is not unduly restrictive in the vast majority of electoral experiments.

**Other causes of election outcomes:** There are arguably many causes of election outcomes. Why should we care so much about one potential cause—an electoral experiment—when other



causes might be more influential (i.e., the economy) or more normatively concerning (i.e., sports game results)? The critical distinction between the experiment and “everything else” is that the experiment constitutes *research*, and is thus subject to research ethics.<sup>9</sup> In the process of doing research, researchers have a responsibility to protect subjects and their communities from possible social harms. Such responsibilities do not extend to non-research activities.

### **8.3 Minimizing aggregate electoral impact as a default**

These arguments for more and less stringent guidance on electoral experimentation suggest that the ethical objective that I elaborate—to avoid changing aggregate election outcomes—represents an intermediate level of scrutiny of these interventions. I argue that this represents an ideal default for experiments, that can be implemented through the framework and decision rule that I advance. However, in any experiment, there will be multiple ethical goals, some of which may come into conflict. When researchers confront conflicting ethical objectives or seek to intervene in environments where their ability to limit aggregate impacts is circumscribed, I echo guidance from American Political Science Association (2020) that exceptions to a principle of minimizing the risk of changing outcomes “should describe plausible impacts at the individual and/or societal level” when justifying intervention.

## **9 Conclusion**

This paper shows that the formalization of an ethical objective can guide researchers to design research consistent with these standards (or avoid research inconsistent with these standards). I find that adherence to ethical goals may come with tradeoffs for learning. Specifically, I show that in electoral experiments, designing experiments with a minimal possibility of changing electoral outcomes can come at a cost to statistical power and external validity.

Similar formalizations of ethical objectives can guide the design of experiments beyond those in elections. Specifically, an application of the framework developed in this article to other experi-

---

<sup>9</sup>I follow the Belmont Report’s definition of research as “an activity designed to test an hypothesis, permit conclusions to be drawn, and thereby to develop or contribute to generalizable knowledge” (p. 3).

mental applications would consist of: (1) a clear mechanism linking the experimental intervention to relevant aggregate/societal outcomes; (2) a maximally agnostic model of how actors' responses to the intervention generate those outcomes; and (3) a set of assumptions restricting the set of actors that might respond to the treatment (the interference assumptions). Within this formulation, elections serve as an "easy" case because they represent a fixed, known mechanism for generating aggregate outcomes. In particular, knowledge of this mechanism increases our understanding of what aggregate impacts are possible and how these impacts are generated. On the other hand, in elections, aggregate outcomes are particularly concerning because the set of impacted actors (residents of a district) is often very large relative to the set of experimental subjects. Development of similar frameworks for community-targeted development projects and audit/correspondence experiments represent important applications that are widespread in the literature (McDermott and Hatemi, 2020).

In sum, I show that careful research design can allow researchers to continue to draw some insights from the experimental study of elections while providing more protections to the communities that we study. This paper advocates a widescale incorporation of ethical considerations as a more prominent guide to research design than is current practice.

## References

- American Political Science Association. 2020. "Principles and Guidance for Human Subjects Research." Available at <https://tinyurl.com/y5vm6cem>.
- Arias, Eric, Horacio Larreguy, John Marshall, and Pablo Querubin. 2019. "Priors Rule: When do Malfeasance Revelations Help or Hurt Incumbent Parties?" Available at [https://scholar.harvard.edu/files/jmarshall/files/mexico\\_accountability\\_experiment\\_v13.pdf](https://scholar.harvard.edu/files/jmarshall/files/mexico_accountability_experiment_v13.pdf).
- Baele, Stéphane J. 2013. "The ethics of New Development Economics: is the Experimental Approach to Development Economics morally wrong?" *Journal of Philosophical Economics* 7 (1): 2–42.
- Beerbohm, Eric, Ryan Davis, and Adam Kern. 2020. "The Democratic Limits of Political Experiments." *Politics, Philosophy, and Economics* 19 (4): 321–342.
- Boas, Taylor, F. Daniel Hidalgo, and Marcus André Melo. 2019. "Norms versus Action: Why Voters Fail to Sanction Malfeasance in Brazil." *American Journal of Political Science* 63 (2): 385–400.

- Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489: 295–298.
- Carlson, Elizabeth. 2020. "Field Experiments and Behavioral Theories: Science and Ethics." *PS Political Science and Politics* (53): 1.
- Catalinac, Amy, Bruce Bueno de Mesquita, and Alastair Smith. 2020. "A Tournament Theory of Pork Barrel Politics: The Case of Japan." *Comparative Political Studies* Forthcoming.
- Davis, Justine, and Kristin Michelitch. 2021. "Field Experiments: Thinking Through Identity and Positionality." *PS Political Science and Politics* Forthcoming.
- Desposato, Scott. 2016. "Conclusion." In *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*. New York: Taylor and Francis pp. 267–289.
- Desposato, Scott. 2018. "Subjects and Scholars' Views on the Ethics of Political Science Field Experiments." *Perspectives on Politics* 16 (3): 739–750.
- Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Gareth Nellis, eds. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. New York: Cambridge University Press.
- Enríquez, José Ramón, Horacio Larreguy, John Marshall, and Alberto Simpser. 2019. "Information saturation and electoral accountability: Experimental evidence from Facebook in Mexico." Working paper.
- Folke, Olle. 2014. "Shades of Brown and Green: Party Effects in Proportional Election Systems." *Journal of the European Economic Association* 12 (5): 1361–1395.
- Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94 (3): 653–663.
- Giné, Xavier, and Ghazala Mansuri. 2018. "Together We Will: Experimental Evidence on Female Voting Behavior in Pakistan." *American Economic Journal: Applied Economics* 10 (1): 207–235.
- Gosnell, Harold F. 1926. "An Experiment in the Stimulation of Voting." *American Political Science Review* 20 (4): 869–874.
- Green, Donald P., and Alan S. Gerber. 2015. *Get out the Vote*. Washington DC: Brookings Institution Press.
- Gubler, Joshua R., and Joel S. Selway. 2016. "Considering the Political Consequences of Comparative Politics Experiments." In *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*, ed. Scott Desposato. New York: Routledge.
- Humphreys, Macartan. 2015. "Reflections on the Ethics of Social Experimentation." *Journal of Globalization and Development* 6 (1): 87–112.

- Hyde, Susan D., and David W. Nickerson. 2016. "Conducting Research with NGOs: Relevant Counterfactuals from the Perspective of Subjects." In *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*. New York: Taylor and Francis pp. 198–216.
- Ichino, Nahomi, and Matthias Schündeln. 2012. "Deterring or Displacing Electoral Irregularities? Spillover Effects of Observers in a Randomized Field Experiment in Ghana." *Journal of Politics* 84 (1): 292–307.
- Kendall, Chad, Tommaso Nannicini, and Francesco Trebbi. 2015. "How do Voters Respond to Information? Information from a Randomized Campaign." *American Economic Review* 105 (1): 322–353.
- Lee, David. 2008. "Randomized experiments from non-random selection in U.S. House elections." *Journal of Econometrics* 142: 675–697.
- Lindbeck, Assar, and Jörgen W. Weibull. 1987. "Balanced-Budget Redistribution as the Outcome of Political Competition." *Public Choice* 52 (3): 273–297.
- López-Moctezuma, Gabriel, Leornard Wantchekon, Daniel Rubenson, Thomas Fujiwara, and Cecilia Pe Lero. 2021. "Policy Deliberation and Voter Persuasion: Experimental Evidence from an Election in the Philippines." *American Journal of Political Science* Forthcoming.
- Manski, Charles E. 2003. *Partial Identification of Probability Distributions*. New York: Springer.
- McDermott, Rose, and Peter K. Hatemi. 2020. "Ethics in Field Experimentation: A Call to Establish New Standards to Protect the Public from Unwanted Manipulation and Real Harms." *Proceedings of the National Academy of Sciences* 117 (48): 30014–30021.
- Michelson, Melissa R. 2016. "The Risk of Over-Reliance on the Institutional Review Board: An Approved Project is Not Always an Ethical Project." *PS Political Science and Politics* April: 299–303.
- Morris, G. Elliott. 2018. "2018 U.S. House Midterm Elections Forecast." Available at <https://www.thecrosstab.com/project/2018-midterms-forecast/>.
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. The Belmont Report: Ethical Principles and guidelines for the protection of human subjects of research. Technical report U.S. Department of Health and Human Services.
- Phillips, Trisha. 2021. "Ethics of Field Experiments." *Annual Review of Political Science* 24: 14.1–14.24.
- Pons, Vincent. 2018. "Will a Five-Minute Discussion Change Your Mind? A Countrywide Experiment on Voter Choice in France." *American Economic Review* 108 (6): 1322–1363.
- Shineman, Victoria. 2020. "Restoring Rights, Restoring Trust: Evidence that Reversing Felony Disenfranchisement Penalties Increases Both Trust and Cooperation with Government." Working paper, available at [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3272694](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3272694).

- Simpser, Alberto. 2013. *Why Governments and Parties Manipulate Elections: Theory, Practice, and Implications*. New York: Cambridge University Press.
- Sinclair, Betsy, Margaret McConnell, and Donald P. Green. 2012. "Detecting Spillover Effects: Design and Analysis of Multilevel Experiments." *American Journal of Political Science* 56: 1055–1069.
- Sircar, Neelanjan, and Simon Chauchard. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Number 10 New York: Cambridge University Press chapter Dilemmas and Challenges of Citizen Information Campaigns: Lessons from a Failed Experiment in India, pp. 287–311.
- Teele, Dawn Langan. 2013. *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*. New Haven: Yale University Press chapter Reflections on the Ethics of Field Experiments, pp. 67–80.
- Teele, Dawn Langan. 2019. "Virtual Consent: The Bronze Standard for Experimental Ethics." In preparation for *Advances in Experimental Methodology* volume.
- Whitfield, Gregory. 2019. "Toward a Separate Ethics of Political Field Experiments." *Political Research Quarterly* (1-12).
- Zimmerman, Brigitte. 2016. "Information and Power: Ethical Considerations of Political Information Experiments." In *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*, ed. Scott Desposato. New York: Taylor and Francis pp. 183–197.