The Ethics of Electoral Experimentation: Design-Based Recommendations

Tara Slough[*]

October 26, 2020

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. 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, 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 elections is that by manipulating characteristics of campaigns, candidates, or voter information, researchers may also be changing aggregate election outcomes.

Two notable changes since the pioneering experimental studies of elections by Gosnell (1926), Eldersveld (1956), Blydenburgh (1971), and Gerber and Green (1999, 2000) influence these ethical considerations. First, researchers now work in contexts with greater variation in electoral institutions and voting behavior than early studies of local elections in US college towns.[1] To the extent that subjects and other residents of their districts internalize the consequences of election outcomes, intervening in different types of elections presents different levels of risks to subjects. Second, the scale of electoral interventions, measured in terms of the number of treated voters, has increased precipitously since early experiments. In addition to academic researchers, campaigns and technology companies now regularly implement massive experimental interventions in elections (see, for example, Pons, 2018; Bond et al., 2012).

I focus on the ethical concern that experimental manipulations may alter aggregate election outcomes. This concern is not new. For example, Dunning et al. (2019) 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" (52). Recently-adopted American Political Science Association (APSA) "Principles and Guidance for Human Subjects Research" echo this concern writing 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). Yet, in a survey of existing experiments, this consideration

[1] Early (pre-2000) experiments occurred in local elections in jurisdictions where researchers worked, namely Chicago, Ann Arbor, and New Haven.

appears to be invoked informally, if at all, in most *ex-post* analyses of electoral interventions. This article proposes a formal, design-based approach to the *ex-ante* consideration of how experimental interventions could affect aggregate election outcomes.

The ethical considerations related to experimental research on elections are admittedly far more complex than the focus on aggregate electoral outcomes in this paper. Notably, Humphreys (2015), Desposato (2018) and Teele (2019) raise questions about standards for consent in field experiments including those on elections. Consent issues represent a distinct ethical consideration from those that I discuss. However, this paper develops other themes articulated in recent literature. Specifically, Beerbohm, Davis, and Kern (2017) argue that experimentation in elections may undermine political equality. Carlson (2020) cautions that researchers must weigh the (possibly overstated) learning benefits of experiments against the potential for harm to subjects.

I provide a new justification for efforts to avoid changing electoral outcomes by focusing on two features of the electoral context. First, contested elections imply an empty Pareto set: changing who wins office harms some individuals while benefiting others. Second, these harms and benefits are distributed across an electoral district, not simply experimental units, as a consequence of the aggregation of votes to the district level. As such, arguments claiming that interventions are welfare enhancing within the subset of district voters in experimental samples generally cannot be informative about the welfare consequences of interventions across the population of plausibly affected individuals. I thus argue that researchers planning experiments should aim to minimize the possibility of changing aggregate electoral outcomes or who wins office.

Minimizing the possibility of changing aggregate electoral outcomes requires a departure from standard practice in the design and analysis of experiments in two ways. 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. As I demonstrate, the frequent omission of information about the relationship between the electoral district and the 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. Moreover, consideration of the aggregation of votes introduces new concerns about researcher culpability for electoral impacts when working with partner organizations (campaigns, NGOS, or interest groups).

Second, while experiments are powerful tools for estimating various forms of *average* causal effects, the 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. Moreover, 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 to 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 least likely to change who wins office.

In response to these concerns, 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 designing an experiment, 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 identify novel trade-offs between ethical considerations and various forms of learning from electoral experiments. By treating fewer voters (all else equal), this ethical consideration admits a trade-off between aggregate electoral impact and statistical power. The concern is particularly acute in cluster-randomized experiments. Avoiding close races implies a trade-off between these ethical principles and external validity–a new source of concern about a common experimental critique. Finally, these guidelines characterize electoral experimentation as a tool that is suitable in some contexts and for some interventions but not others.

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. 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. Because these assumptions must be invoked *ex-ante*, more careful consideration of possible general equilibrium effects is critical for *ex-ante* consideration of the ethical implications of an experiment.

This paper makes three contributions. First, it develops tools to guide researchers considering prospective interventions on elections, as well as consumers of research describing such interventions. I show how these considerations depart from current practices in the reporting of electoral experiments. Further, I illustrate the utility of these tools on electoral data from the US, simulating admissable experimental designs under the decision rule advocated. Second, I identify a set of trade-offs inherent to the design of electoral experiments that emerge in the consideration of whether experiments change electoral outcomes. Characterization of these trade-offs allows for a richer discussion about the merits and limitations of experiments on elections as as a research design for learning about political behavior, persuasion, and electoral accountability. Finally, I advance the view that ethical considerations should be paramount when designing experimental interventions. While some existing works adopt non-standard experimental designs as a function of ethical considerations (i.e., Slough and Fariss, 2020), to my knowledge, this is the first general framework to incorporate ethical concerns across a range of experimental designs in a common setting (elections). Such a framework may inform the development of other design-based strategies to reduce ethical concerns in other social science experiments.

2 Defining the Ethical Objective

Intervening in elections presents risks for precisely the reasons that we study elections: because "elections have consequences" for governance, policymaking, and welfare.[2] In principle, such consequences constitute a basis for the set of possible harms and benefits to subjects. In considering these outcomes, the electoral setting is unique among other field experiments because elections generate winners and losers through a fixed (known) aggregation mechanism.

In contested elections, the set of possible Pareto-improving interventions is generally empty: an intervention will harm some subject such as a candidate made to lose support while accruing benefits to another subject such as a candidate with increased support. Even if we were to restrict the subject designation to voters – exempting candidates or public officials – so long as preferences vary across the electorate, the possibility of shifting support from one candidate to another ostensibly generates harms and benefits to different groups of voters. In this sense, we know that electoral interventions can generally harm some actors.

The aggregation of votes to determine outcomes presents normative considerations unique to the electoral context. Importantly, it suggests that the set of individuals that realize the consequences of an intervention includes members of a district, which often far surpasses the subset of registered voters considered experimental "subjects." This consideration weakens the merits of some standard efforts to mitigate harm to subjects in experiments. In other (heavy touch) field experiments that could generate harm to subjects, researchers often purport to have randomized an intervention that would happen anyway (i.e., by a government or aid group), in so doing gleaning epistemic benefits without imparting additional "harm" (Teele, 2013). However, differences between non-experimental and experimental allocations of an electoral intervention can lead to very different distributional outcomes (of harm and benefit) as a result of aggregation of votes. Further-

[2] Indeed, downstream analyses of electoral experiments do suggest that who wins office (or how office is won) is consequential for later policymaking (Ofosu, 2019; Gulzar and Khan, 2018).

more, claims about welfare among experimental subjects (however operationalized) are generally uninformative about the welfare of the full pool of individuals that realize electoral consequences.

Some experimental interventions like anti-vote buying campaigns or revelation of corruption/malfeasance to voters serve as hypothesized antidotes to bad governance. In the case of a malfeasant incumbent, there may be a motivation to *change* aggregate electoral outcomes, even if it harms the malfeasant candidate, her cronies, and her clients. This motivation relies on an assumption that welfare would improve if a different candidate were elected as consequence of the intervention. In so doing, it imposes an unspecified welfare criterion in addition to assumptions about the features of the election, e.g., that the challenger pool includes non-malfeasant candidates. The strength of these assumptions presents a tension with the use of an electoral experiment. Specifically, if a researcher were well-positioned to make these assumptions, there is presumably less to be learned from an experiment, and possibly even an argument against withholding a presumed welfare-enhancing intervention from control-group voters. I contend that in these assumptions are too strong for most electoral settings. As such, researchers should design experiments to avoid changing aggregate electoral outcomes.

2.1 Experiments and their Counterfactuals

A focus on the effects of experimental interventions on aggregate election outcomes requires consideration of what would happen in the absence of the experiment. The interventions that are randomized in an experiment may or may not occur absent the experiment. The relevant consideration is thus: how could the experimental allocation of the intervention change electoral outcomes? Because consequences emerge from the aggregation of votes, the impact of changing (randomizing) the allocation of an intervention depends critically on the mapping between: (i) the unit and saturation of treatment allocation in the experiment; (ii) the unit and density of treatment allocation in the non-experimental implementation (if one exists); and (iii) the relationship between the treated units and the electoral district. Given these considerations, it is important to specify concretely the counterfactual in the absence of the experiment.

To this end, I consider two cases of electoral experiments in Table 1. Consider the modal first

	Actors				
Cases	Researcher	Partner	Experiment	Counterfactual (absent experiment)	Example
1	✓		*Researcher designs, implements experimental intervention.* (Note: An partner may participate in or endorse the experiment, but experiment is initiated by researcher and intervention is funded through the researcher or externally.)	*No intervention occurs.*	Experiments conducted in Metaketa-I and documented in Dunning et al. (2019)
2	✓	✓	*Researcher randomizes a partner-funded and implemented intervention.*	*Partner funds and implements intervention without randomizing allocation of treatment, possibly with less data collection.*	Pons (2018)

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

case (Case #1) of electoral experiments in which a researcher designs and implements an intervention that would otherwise not have occurred. A researcher values learning; the most common quantitative operationalization of learning in experimental design is statistical power.[3] Given that a researcher cannot control how subjects will respond to a treatment or know *ex-ante* the precision gains from blocking or covariate adjustment, she typically maximizes power subject to a budget or implementation constraint by increasing the number of subjects in an experiment. Yet, increasing the number of subjects increases the range of possible electoral impacts either directly through treated voters (Section 4) or through general-equilibrium responses (Section 5).

Now consider the case in which some intervention by an partner – usually an NGO or interest group – is modified to include an experimental component (Case #2). The inclusion of an experiment generally changes allocation of the intervention. Since the relevant unit of aggregation is the electoral district, consider three possible changes in the allocation treatment from the non-experimental case. First and closest to the first case, the intervention may be implemented in districts that where it would otherwise not have been implemented in the non-experimental regime. For example, in a cluster-randomized experiment, in search of statistical power, researchers will generally look to expand the number of clusters (as opposed to individuals per cluster), which may

[3] If learning consists of Bayesian updating, a measure of difference between prior and posterior may be a better operationalization of learning. Given that learning consists of both a possible change in the mean and dispersion of beliefs, a better-powered designs roughly approximate learning in terms of a reduction in posterior variance.

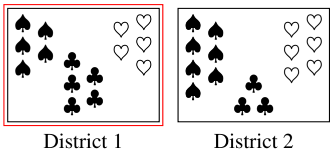
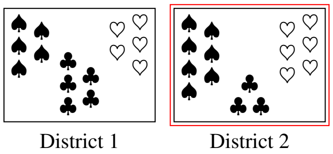
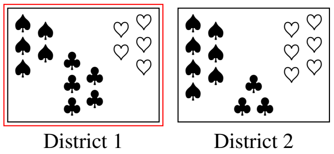
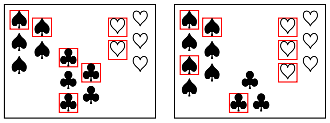
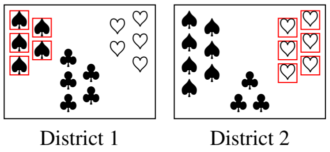
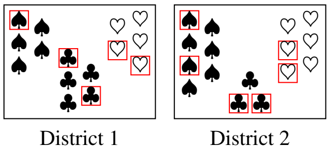
	Counterfactual allocation	Experimental allocation	Description
1			Intervention is implemented in **different districts** under experimental allocation.
2			Intervention is implemented at **different saturation** under experimental allocation.
3			Intervention is targeted to **different voters** under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

Table 2: Illustration of three ways the allocation of an intervention by a partner can be changed when allocated experimentally. Each symbol represents a voter. The red boxes indicate the districts or individuals assigned to treatment.

expand the intervention into additional or different districts. Second and generalizing this point, there may be a change in the proportion of a district that is treated (differential saturation). This may stem from increased implementation costs for delivering a randomly-assigned treatment or simply the need for control units. Finally, it may be the case that different voters in a district are treated under an experimental allocation. For example, if a campaign targets its message all probable swing voters, under an experimental allocation, such voters must be assigned to control with some positive probability. Even holding constant the number of treated voters in a district, the types (and thus voting behavior) of treated voters may change when incorporating the experimental allocation of treatment. Table 2 visualizes these three changes in the allocation of treatment due to the use of an experiment.

Ultimately, I argue that ethical concerns about an experiment changing aggregate electoral outcomes must 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. Within the framework I advance, a researcher collaborating with a partner need not be "held responsible" for the impact of electoral interventions that would target the same voters under the experiment and its counterfactual. At the same time, collaboration with a partner does not absolve a researcher from considerations about how the experiment changes allocation of the treatment within or accross districts. Importantly, analysis of the difference between the experimental and non-experimental allocation of an electoral intervention cannot be completed without reference to the district or unit of aggregation.

2.2 The Ethical Objective

I assume that researchers' ethical objective is to avoid changing who ultimately wins office, relative to what would have happened absent the experimental allocation of an electoral intervention. In the aggregate, thus, 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 (no mandate effects).

Some existing 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; Cox and McCubbins, 1986; Smith and Bueno de Mesquita, 2011; 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 subsequent 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 approach that I advocate considers two types of uncertainty that we have as researchers. First, as elaborated above, we lack the ability to determine the welfare effects of an intervention or even a common welfare criterion upon which such a determination could be made. Indeed, any electoral outcome is apt to produce winners and losers. The approach here simply asserts that researchers should not be determining who wins and who loses in the service of research. Second, 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 and compared to predictive distributions characterizing relevant measures of closeness in elections.

When does an elections experiment become unacceptable on grounds that it is too likely to change election outcomes? In principle, we could eliminate the risk of influencing electoral outcomes entirely by not running these experiments. Yet, we also learn about political behavior, persuasion, and electoral accountability from these interventions. Some existing experimental interventions are small (or sparse) enough to have a near-negligible effect on electoral outcomes, even by the conservative standards specified in this article. This article provides a systematic way to bound possible effects *ex-ante*. It then suggests ways to compare these bounds to predicted outcomes in order to determine how 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 the 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 critical 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.

Researchers define the clustering of subjects within a district. I assume that voters in district d, indexed by $j \in \{1, ..., n_d\}$ 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 $\sum_c n_c = n_d$. In service of generality, 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, I remain agnostic as to the (generally non-random) allocation mechanism. This notation allows for characterization of four principal strata, described in Table 3. I use the notation S^{cd}_{11}, S^{cd}_{10}, S^{cd}_{01}, and S^{cd}_{00} 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^{cd}_{11}	Always assigned	1	1	$\lvert S^{cd}_{11}\rvert = 0$	$\lvert S^{cd}_{11}\rvert \geq 0$
S^{cd}_{10}	If-experiment assigned	1	0	$\lvert S^{cd}_{10}\rvert > 0$	$\lvert S^{cd}_{10}\rvert \geq 0$
S^{cd}_{01}	If non-experiment assigned	0	1	$\lvert S^{cd}_{01}\rvert = 0$	$\lvert S^{cd}_{01}\rvert \geq 0$
S^{cd}_{00}	Never assigned	0	0	$\lvert S^{cd}_{00}\rvert > 0$	$\lvert S^{cd}_{00}\rvert \geq 0$

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^{cd}_{10} – individuals exposed to the treatment because it is assigned experimentally – and S^{cd}_{01} – 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:

$$\mathcal{S}_d = \frac{\sum_{c \in d} |S^{cd}_{10} \cup S^{cd}_{01}|}{n_d} \tag{1}$$

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.[4] 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

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

literature, I assume the potential outcomes framework as 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.).

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 (Manski, 2003). In invoking these bounds, I make no assumption about the plausible effects of an intervention (i.e., monotonicity). Indeed, Bayesian models of voter updating invoked in informational electoral experiments predict non-monotonicity in treatment effects as a function of the location of the signal relative to the prior (i.e. "good news" 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 not appropriate, but the bound on aggregate electoral impact will 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:*

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

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 "margin to pivotality," ψ_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, this is the margin of victory. In a proportional representation (PR) system, there are various interpretations of ψ_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 A 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 (outcomes). 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 ψ_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 ψ_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 ψ_d, $\widehat{f}(\psi_d) \sim f(\psi_d|\widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}}$ 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 ψ_d. In particular, I suggest that researchers calculate a threshold $\underline{\psi_d}$, that satisfies $\widehat{F}^{-1}(0.05) = \underline{\psi_d}$, where $\widehat{F}^{-1}(\cdot)$ indicates the quantile function of the predictive distribution of ψ_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 ψ_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 in some part of the world than others. Both implications of the decision rule posit implications for the external validity of electoral experiments and the (non)-universal applicability of electoral experiments as a tool, points to which I return in Section 7.

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 (2017) about electoral experiments generating various forms of political inequality.

5 When is this Analysis Non-Conservative?

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 obtain, this analysis may 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 voters 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:

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

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.[5] In order to account for between-cluster interference, a

[5] Note that identification of causal estimands is not the concern here. The concern is that some manifestation of the treatment (or response to the treatment) could alter the votes of a growing portion of a district.

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.

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

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, $MAEI_d^{wb} \geq MAEI_d^{w}$.

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:

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

However, it must always be case that the margin to pivotality, $\psi_d \leq \frac{1}{n_d}\sum_{c\in d}E[a_c(0)]n_c$, 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 π_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 π_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 acquaintances/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.[6]

To the extent that scholars have 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

[6] This calculation is based on report of 100,000 flyers in Willis (2014).

$\pi_c > 0$, 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 (whether in the same district or otherwise), the bound widens. Most challengingly, such a determination must be made before the intervention is fielded.

6 Illustration: Existing Experiments and Simulation

I now consider how the framework described here can be employed in the planning of an electoral experiment. I first provide an overview of how the framework can be applied to existing studies collected by Enríquez et al. (2019). This exercise reveals substantial variability in the estimated $MAEI_d$'s. It also suggests that the framework is most logically (and productively) applied *ex-ante* (before a study goes to the field) rather than *ex-post* (in the analysis of experimental data). To this end, I simulate a series of experimental research designs using real administrative data that speak to the *ex-ante* application of this framework.

6.1 Relation to Electoral Experiments on Information about Incumbent Performance

I focus on back-of-the-envelope calculation of the $MAEI_d$ given information reported in articles and appendices only. I use back-of-the-envelope calculation as opposed to consulting replication data for two reasons. First, these calculations survey whether information necessary to (begin to) aggregate votes is reported and what barriers to these calculations exist. Second and more practically, many of these studies are still unpublished, rendering justifiably limited access to replication data. I report the studies, their relationship to the proposed framework, and the calculations executed in Appendix Table A1. I lack any *ex-ante* information about how to predict these races, so I focus only on the calculation of $MAEI_d$ under Assumptions 1 and 2.

Thirteen of the 14 studies intervene in multiple races (districts). I focus on calculating either an *average* $MAEI_d$ across districts. The *average* $MAEI_d$ is an abstraction from the decision rule described in this paper. However, for the purposes of examining the literature, it does serve as a measure of the variability across studies on this metric. I am only able to estimate the $MAEI_d$

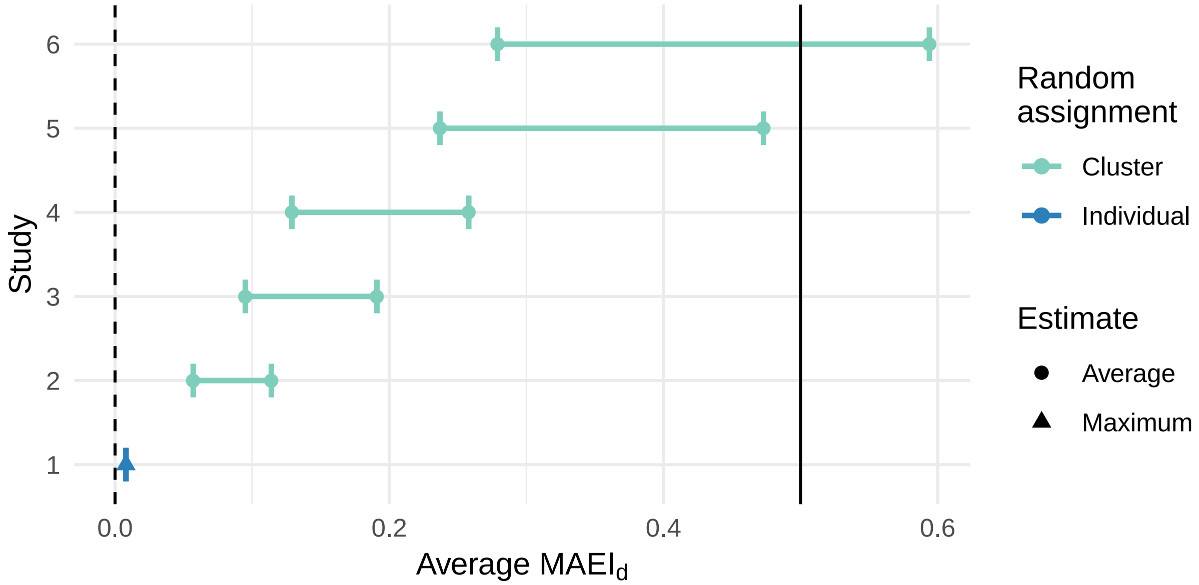


Figure 1: Estimated maximum or average $MAEI_d$ for six electoral experiments on electoral accountability. The interval estimates in the cluster-randomized experiments indicate the range of $MAEI_d$ estimates for any $E[a_c(0)] = [0.5, 1]$. For discussion of these calculations, see Table A1.

in six of 14 studies, varying $E[a_c(0)]$ from its minimum of 0.5 (for all c) to its maximum of 1 (for all c). I present these estimates in Figure 1. The graph suggests that the maximum degree to which existing experiments could have moved electoral outcomes varies widely. Recall that these estimates in isolation cannot assess whether an intervention was consistent with the decision rule advocated here because I lack data on the predicted margin of victory. Nevertheless, any $MAEI_d > 0.5$ can never pass the decision rule, regardless of the predicted margin to pivotality. One immediate concern from Figure A1 is that cluster assigned treatments appear to be assigned at a very high density within districts.

The barriers to estimation of the $MAEI_d$ in the remaining eight studies are informative for how we think of electoral impact. In general, these studies do not provide information on how the experimental units relate (quantitatively) to the electorate as a whole. 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). Note that the non-random sampling is generally well-justified from a design perspective and the constituency size is not necessary for the estimation of causal effects. The takeaway from this survey of 14 studies is simply that considerations of aggregate electoral impact require analyses that are not (yet) standard practice. The variation in Figure 1 suggests that research designs vary substantially on this dimension and justify these considerations.

6.2 Simulations Using Electoral Data

I conduct a simulation of the proposed guidelines for research design with electoral data from the US state of Colorado. The purpose of the simulation is to demonstrate how the framework proposed here can be used in practice and illustrate insights from the model. In the simulation, I rely on real voter registration data, precinct-to-district mappings, and election predictions. I manipulate aspects of the experimental design, particularly the method of assignment to treatment (clustered or individual), the number of units assigned to treatment, and the allocation of treatments across districts.

Specifically, I simulate a hypothetical experiments to be implemented in 2018 elections in Colorado. Because elections are administered at the state level in the United States, the simulations are greatly simplified by focusing on a single state. Moreover, all races in 2018 were at the state level or below. Colorado was randomly selected, though this draw was "lucky" for two reasons. First, Colorado exhibits substantial geographic variation in the distribution of political preferences, allowing examination of these design principles over a heterogenous set of districts. Moreover, Colorado provides partisan voter registration data disaggregated to the precinct level which assists in prediction.

In the simulations, I assume that an experimental intervention would not occur absent the researcher (Case #1 above). In the case of Colorado, there are many forecasts available for the 2018 US House elections; I do not know of forecasts for State House seats. Therefore, in the case of the State House races, I predict outcomes from (limited) available data, namely partisan voter registration data and lagged voting outcomes. I train a very basic predictive model on electoral data from 2012-2016 (three elections) and then predict outcomes for 2018.[7] I outline my prediction method

[7] One concern is that this model does not incorporate time shocks (in this case, the "blue wave")

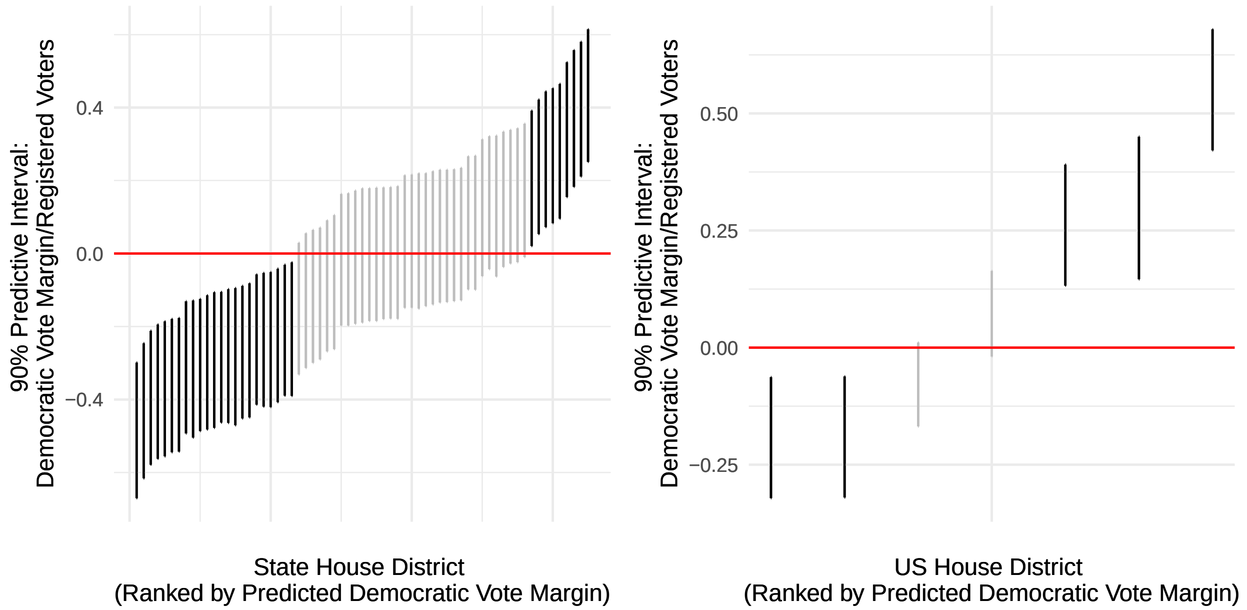


Figure 2: Predictive intervals for 65 State House seats and 7 US House seats in 2018. State House predictions are calculated following the method in Section C.3 while the US House predictions are "off the shelf" from Morris (2018). Grey lines represent grounds for declining to conduct an experiment in a district under the decision rule proposed here.

and the construction of the predictive distribution ($f(\psi_d|\widehat{\boldsymbol{\theta}})$) in Appendix C.3.

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 pass the decision rule proposed in this paper. In sum, 33/65 State House races and 2/7 US House races bound 0. More precise prediction algorithms, particularly in the State House races, may alleviate concerns in some cases. On the other hand, these (effectively) two-canddate races are relatively predictable given the comparative salience of partisanship in the US and vast amount of effort devoted to predicting and understanding US elections. As such, this analysis represents roughly a "best case" scenario for experimental design.

I consider several variants of research designs, each invoking SUTVA and, by design, satisfying Assumption 2.[8] I first consider experiments that assign individual voters (not clusters) to treatment.

absent polling data. More sophisticated predictive algorithms can easily be incorporated.

[8] I assume all voters in cluster-randomized designs are assigned to treatment if they belong to a treated cluster. For individually randomized experiments, Assumption 2 is implied by SUTVA.

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, as would be the case for an experiment on likely Republican (resp. Democratic) voters. 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 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 (in either direction), 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 doubles the number of subjects, as implied by Equation 2. Third, comparing the top to 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 A2 for summary statistics). Finally, when describing the maximum number of treated subjects as a proportion of the electorate, in general, 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.

Moving to a cluster-randomized treatment at the *precinct* level, Figure 4 examines the expected number of precincts that could be assigned to treatment in each type of race. These graphs assume that all registered voters in a treated precinct are treated, or equivalently that Assumption 2 (no within-cluster interference) does not hold for some lower level of treatment saturation. The graph

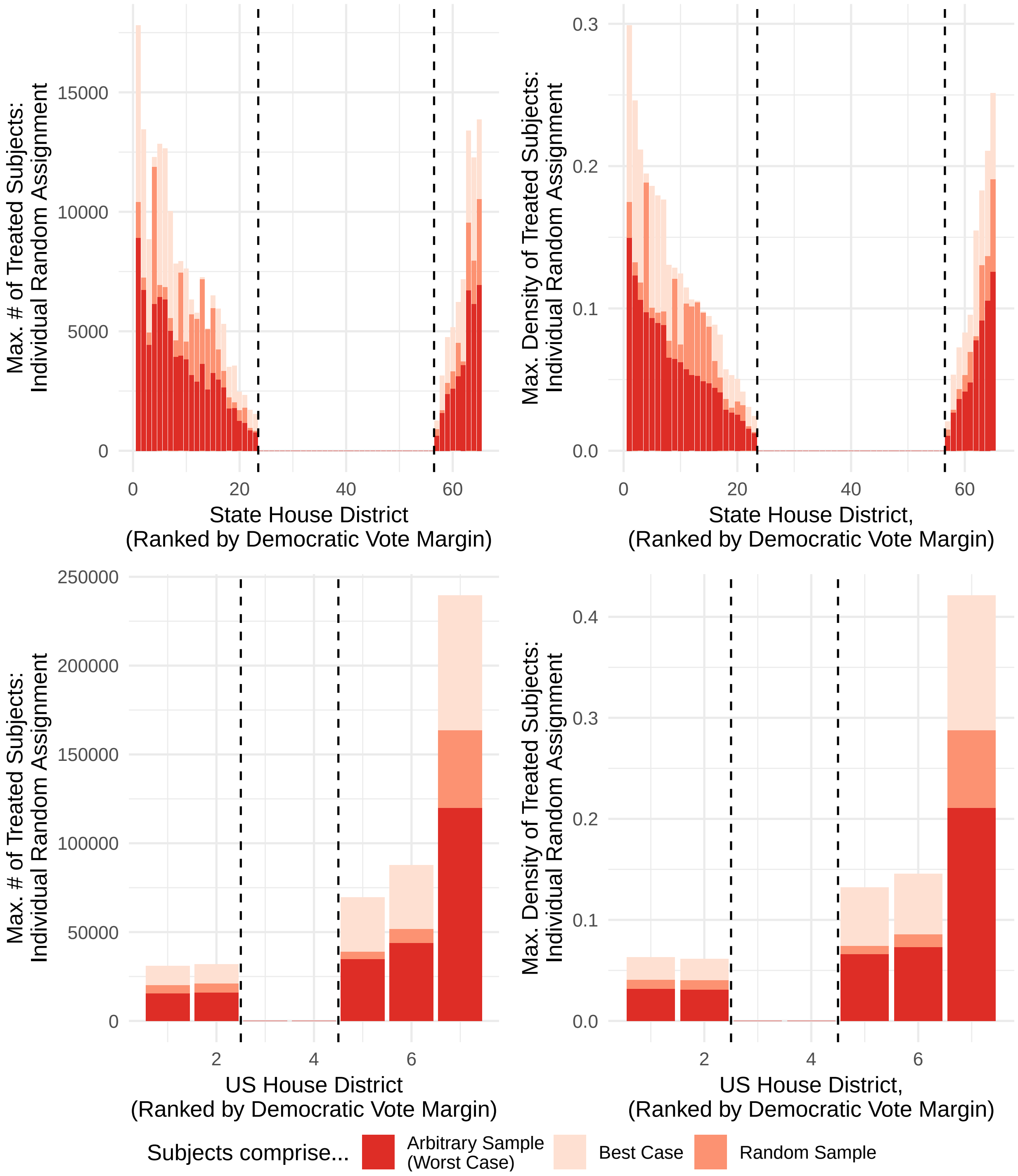


Figure 3: Maximum number of individuals (left) or and individuals as a proportion of registered voters (right) that can be assigned to treatment under decision rule. The black dotted lines denote the districts excluded on the basis of predictions bounding 0.

suggests that in State House races, the number of "treatable" precincts is quite small. Given that cluster-randomized electoral treatments are most popular in low-level races in developing settings, this suggests that researchers should be much more cautious about treating large segments of a district, even in "best case" scenarios elections are relatively predictable. In general, if researchers must treat at the cluster level, they should minimize the number of clusters treated in any given district.

7 Implications for Research Design and Learning

The parameters used to characterize the possibility for electoral interventions to change 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 on the ethical grounds spelled out in this paper in some contexts than others.

7.1 Electoral Systems, Rules

Electoral systems specify the mapping between votes and seats, influencing the plausible range of ψ_d, the margin to pivotality. Consider the distinction between elections using first past the post (FPTP) majoritarian and a closed list proportional representation (PR). While either type of race can, in principle, be arbitrarily close, the maximum value of ψ_d is given by the reciprocal of district magnitude. In a FPTP election with one office at stake, the theoretical upper bound on ψ_d is 1. In a PR race with standard Hare or D'Hondt seat allocation formulas, the upper bound on ψ_d clearly falls quickly as district magnitude increases. Increases in proportionality under PR thus limit the possibility of "landslide" elections where high-density treatments would be unable to move outcomes.

More subtle variants of majoritarian and PR electoral systems also exaggerate or limit the degree of variation in ψ_d. For example, two-round systems or runoff elections create complications in prediction of ψ_d, at least in advance of a first round. Moreover, if changing electoral outcomes also includes changing which politicians win office (as opposed to which parties win seats), open list PR systems imply that ψ_d can be interpreted in terms of the last seat allocated *or* the allocation

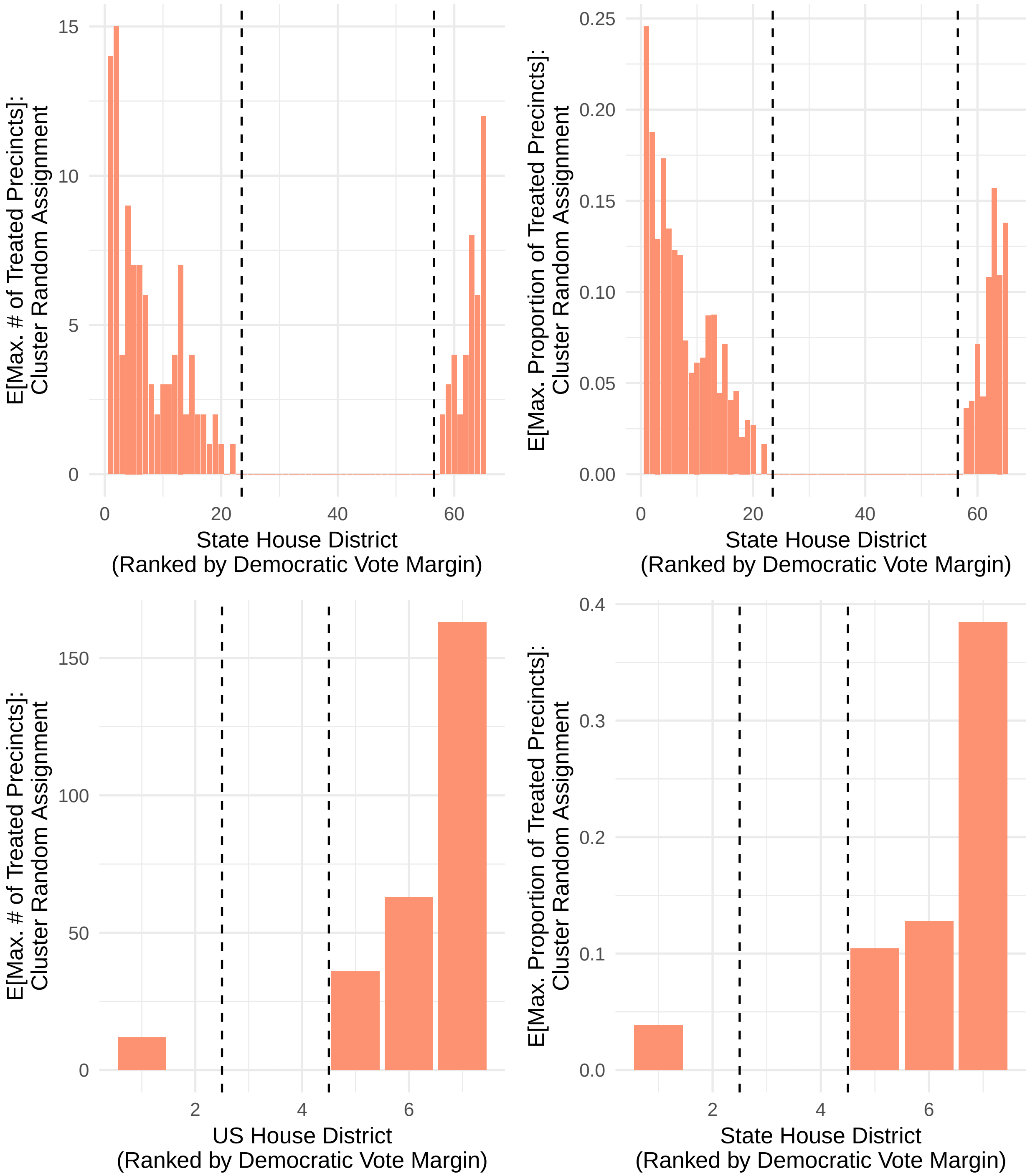


Figure 4: Maximum number of precincts (left) or and precincts as a proportion of all precincts (right) that can be assigned to treatment under decision rule. The expectation is evaluated over simulated assignments of treatment to precincts with heterogeneous numbers of registered voters. Expectations are rounded down to the largest number of full precincts that pass the decision rule. The black dotted lines denote the districts excluded on the basis of predictions bounding 0.

of seats within a list. This can reduce ψ_d, reducing the admissible $MAEI_d$ under the decision rule and complicates prediction of $\widehat{\psi_d}$.

Beyond considerations of the plausible range of ψ_d, there remain questions about our ability to predict this quantity. In FPTP elections, particularly in the context of two parties, ψ_d is easily interpretable as the margin of victory (as a ratio of registered voters). Yet, in other types of elections, ψ_d is a much more abstract and less studied quantity. The extent to which we can predict $\widehat{f}(\psi_d)$ across contexts remains an open empirical question that is quite important for the considerations described here.

Other features of the electoral system may also condition the suitability of a race for experimentation. Consider the role of concurrent elections. Concurrent elections often include races with different electorate sizes, captured by the n_d parameter in the model. An intervention on a set of voters may represent a much larger proportion of the electorate in one race than in a concurrent race. Note that concurrent elections do not represent any form of experimental design violation in standard experimental analyses. In general, concerns of concurrent elections are reduced if the experimental manipulation happens in the "smallest" race: the race with the smallest n_d. However, if the number of voters in the smallest race is greater than $\frac{1}{2}$ the number of voters for a larger race, then variation in competitiveness could compensate for variation in saturation. As this discussion outlines, concurrent elections can lead to profound differences in assessments of the risk of electoral experiments.

7.2 Contextual Features of Elections

We observe substantial variation in our ability to predict elections across contexts. Most predictive models, including those described here leverage some combination of past electoral returns and contemporaneous indicators (polls, incumbency, and macroeconomic indicators). It is possible that in other contexts, predictive models could exploit ethnicity or other identity-based characteristics that are highly prognostic of voter behavior. Yet, there exist three kinds of variation in our ability to precisely estimate the predictive distribution, $\widehat{f}(\psi_d)$, accross contexts. First, the levels of effort invested in prediction of elections vary by country and office. At the national level executive and congressional elections in the U.S. and other OECD democracies, there is substantially less effort devoted to develop prediction methods for lower level offices or elections in developing countries. To the extent that limited effort stymies the precision of prediction efforts, a researcher would be *less* likely to experiment in elections for which predictive models are underdeveloped.

Second, the degree to which voter characteristics or past voter behavior in election $t - 1$ is predictive of their behavior in election t varies substantially. When electoral volatility is high or parties are weakly institutionalized, our ability to identify "safe" districts on the basis of observables is apt to be curtailed.

Finally, the amount of information collected to predict voter behavior varies across contexts. In particular, the availability of polls varies drastically. Absent contemporaneous information, identifying large aggregate shifts (i.e, the "Blue wave" in the 2018 US Congressional elections) becomes more difficult.

7.3 Administrative Data Availability

The availability of administrative electoral returns (as opposed to surveys) often conditions considerations about the unit of clustering. Where we have accurate measures of relevant individual-level outcomes, randomization at the individual level is often preferred to maximize power. However, in most elections, the ballot is secret. While turnout may be observed in administrative records in some countries, where data equivalent to voter files are unavailable or turnout is not the central outcome of interest, researchers are often forced to treat precincts (or the lowest level of administrative electoral return aggregation). Yet, in order to detect any treatment effect on aggregate vote shares, researchers tend to treat a larger share of the precinct, increasing $\mathcal{S}_d$. Thus, by informing the level of clustering, the availability of administrative data also conditions the risk presented by electoral experiments.

7.4 Trade-offs and Implications for Knowledge Cumulation

The above discussion posits five main design choices by which researchers can limit the possibility that their experimental interventions change who wins elections:

1. Reduce the number of voters assigned to treatment.

2. Avoid implementing experimental interventions in close or unpredictable races.

3. Implement interventions in larger electoral districts.

4. Experiment in FPTP races.

5. Select treatments to improve the plausibility of assumptions of restricted interference.

Yet, these design strategies posit trade-offs in terms of learning from electoral experiments. First, consider the implications of #1 for statistical power. Constraints on power, at least in terms of the number of observations, N, typically emerge from inability to treat enough individuals or clusters due to budgetary constraints. This paper holds that power may be further constrained by concerns about minimizing electoral impact when experimenters reduce the density of treatment. This trade-off is particularly salient in experiments seeking to analyze aggregate electoral outcomes at the cluster (i.e. polling station or precinct) level. Thus, I identify a likely tension between the ethical design of electoral experiments and their statistical power.

A further implication of this trade-off between statistical power and the number of voters experimentally assigned to treatment is that researchers should be careful not to "over-power" electoral experiments by including ever-increasing samples of voters. While power is increasing in the number of subjects or clusters (N), as N grows large, the marginal power gains from adding additional subjects is decreasing (for sufficent N). Importantly, the possible electoral impact of an intervention increases linearly in the number of treated units,[9] suggesting that above some threshold, the increased risk of impacting elections outweighs precision gains from increasing the sample size. In a time when interventions are becoming cheaper to implement to large swaths of the electorate via SMS or social media, researchers should justify their sample selection carefully to avoid the possibility of changing electoral outcomes.

[9] This assumes that units are randomly sampled from a larger population, or that $E[a_c]$ remains constant as N increases.

Strategy #2 – avoiding close or unpredictable races – raises a possible trade-off bentween ethical design of electoral experiments and external validity. In light of these ethical considerations, researchers would ideally maximize the "margin to pivotality," or ψ_d. In such races, the ability of an experiment to change who wins office is lower. 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 in terms of the possible risks to election outcomes and the generalizability of insights about behavior. 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.

A focus on landslide races may also affect statistical power. However, the direction of these effects is ambiguous. Following the discussion of external validity, there may be fewer persuadable ("swing") voters in landslide districts than in marginal districts. If treatment effects depend on persuasion, a lack of persuadable voters could reduce 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 magnitude is higher in very imbalanced electorates. The net implications for power are therefore ambiguous.

Strategies #3 and #4 constrain the types of races in which intervention consistent with the ethical guidelines in this article is feasible. These strategies posit rule out electoral experimentation in some countries or offices as a function of the electoral system or institutions. This creates challenges for the development of general or comparative knowledge. Finally, Strategy #5 circumscribes the set of treatments that reseachers 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.

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 treament effect ($LATE$) 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 PR contests (but see, e.g., Folke, 2014; Fiva, Folke, and Sørensen, 2018). If the limitations on the appliation of electoral experiments discussed here 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 be more interesting or 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 wholescale abandonment of the tool.

8 Conclusions

This paper advances two broad insights for future research. First, I show 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. Second, I posit a new justification for the study of the broader societal outcomes of electoral experiments via who wins office. The absence of the mapping between experimental units (e.g., precincts) and the districts at which votes are aggregated in existing studies (Table 1) demonstrates that these considerations are not yet standard practice.

Is the ethical standard of not changing election outcomes a one-size-fits all criterion for electoral experiments? While this criterion is the most widely-invoked ethical consideration about outcomes of interventions, it need not be the only ethical consideration. The formulation in this article provides one way to weigh different ethical commitments. For example, in the experimental study of representation in elections, recruiting different candidates implies a *district*-level treatment, which can never pass the decision rule in this paper. However, there may be other arguments in favor of such interventions that outweigh the considerations I advance. Following 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" (15), this article provides researchers an analytical framework through which to justify the tradeoffs between ethical desiderata.

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 they study. While certain aspects of the present discussion are distinct to the electoral and/or experimental context, further work is needed to understand how ethical goals can inform decisions about research design in other research contexts and what trade-offs these decisions may entail. This paper advocates a widescale incorporation of ethical considerations as a more prominent guide to research design than is current practice.

References

Adida, Claire, Jessica Gottlieb, Eric Kramon, and Gwyneth McClendon. 2017. "Reducing or Reinforcing In-Group Preferences? An Experiment on Information and Ethic Voting." *Quarterly Journal of Political Science* 12 (4): 437–477.

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`.

Banerjee, Abhijit, Selvan Kumar, Rohini Pande, and Felix Su. 2011. "Do Informed Voters Make Better Choices? Experimental Evidence From India." Available at `https://scholar.harvard.edu/files/rpande/files/do_informed_voters_make_better_choices.pdf`.

Beerbohm, Eric, Ryan Davis, and Adam Kern. 2017. "The Democrati Limits of Political Experiments." Working paper, available at `https://scholar.harvard.edu/files/beerbohm/files/democratic_limits_of_political_experiments_eb_rd_ak.pdf`.

Bhandari, Abhit, Horacio Larreguy, and John Marshall. 2020. "Able and Mostly Willing: An Empirical Anatomy of Information's Effect on Voter-Driven Accountability in Senegal." *American Journal of Political Science* Forthcoming. Available at `https://scholar.harvard.edu/files/jmarshall/files/accountability_senegal_paper_v5.pdf`.

Blydenburgh, John C. 1971. "A Controlled Experiment to Measure the Effects of Personal Contact Campaigning." *Midwest Journal of Political Science* 15 (2): 365–381.

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* forthcoming.

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.

Buntaine, Mark T., Ryan Jablonski, Daniel L. Nielson, and Paula M. Pickering. 2018. "SMS Texts on Corruption Help Ugandan Voters Hold Elected Councillors Accountable at the Polls." *Proceedings of the National Academy of Sciences* 115 (26): 6668–6673.

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.

Chong, Alberto, Ana de la O, Dean Karlan, and Leonard Wantchekon. 2015. "Does Corruption Information Inspire the Fight or Quash the Hope? A Field Experiment in Mexico on Voter Turnout, Choice, and Party Identification." *Journal of Politics* 77 (1): 51–77.

Cox, Gary W., and Matthew D. McCubbins. 1986. "Electoral Politics as a Redistributive Game." *The Journal of Politics* 48 (2): 370–389.

Cruz, Cesi, Philip Keefer, and Julien Labonne. 2018. "Buying Informed Voters: New Effects of Information on Voters and Candidates." Available at https://static1.squarespace.com/static/58c979fad1758e09d030809c/t/5c048e82898583120b1f73cc/1543802523246/buying_informed_voters_web.pdf.

Cruz, Cesi, Philip Keefer, Julien Labonne, and Francesco Trebbi. 2019. "Making Policies Matter: Voter Responses to Campaign Promises." Working paper available at https://static1.squarespace.com/static/58c979fad1758e09d030809c/t/5cfed616d6104500019dff1b/1560204824899/making_promises_matter_6102019.pdf.

de Figueiredo, Miguel F.P., F. Daniel Hidalgo, and Yuri Kasahara. 2011. "When Do Voters Punish Corrupt Politicians? Experimental Evidence from Brazil." Available at https://law.utexas.edu/wp-content/uploads/sites/25/figueiredo_when_do_voters_punish.pdf.

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.

Eldersveld, Samuel J. 1956. "Experimental Propaganda Techniques and Voting Behavior." *American Journal of Political Science* 50 (1): 154–165.

Enríquez, José Ramón, Horacio Larreguy, John Marshall, and Alberto Simpser. 2019. "Information saturation and electoral accountability: Experimental evidence from Facebook in Mexico." Working paper.

Fiva, Jon H., Olle Folke, and Rune J. Sørensen. 2018. "The Power of Parties: Evidence from Close Municipal Elections in Norway." *The Scandanavian Journal of Economics* 120 (1): 3–30.

Folke, Olle. 2014. "Shades of Brown and Green: Party Effects in Proportional Election Systems." *Journal of the European Economic Association* 12 (5): 1361–1395.

George, Siddharth, Sarika Gupta, and Yusuf Neggers. 2018. "Coordinating Voters against Criminal Politicians: Evidence from a Mobile Experiment in India." Available at https://scholar.harvard.edu/files/siddharthgeorge/files/voter_mobile_experiment_181126.pdf.

Gerber, Alan S., and Donald P. Green. 1999. "Does Canvassing Increase Voter Turnout? A Field Experiment." *Proceedings of the National Academy of Sciences* 96 (14): 10939–10942.

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.

Gulzar, Saad, and Muhammad Yasir Khan. 2018. "Motivating Political Candidacy and Performance: Experimental Evidence from Pakistan." Working paper.

Humphreys, Macartan. 2015. "Reflections on the Ethics of Social Experimentation." *Journal of Globalization and Development* 6 (1): 87–112.

Humphreys, Macartan, and Jeremy M. Weinstein. 2012. "Policing Politicians: Citizen Empowerment and Political Accountability in Uganda - Preliminary Analysis." IGC Working Paper S-5021-UGA-1.

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.

Lee, David. 2008. "Randomized experiments from non-random selection in U.S. House elections." *Journal of Econometrics* 142: 675–697.

Lierl, Malte, and Marcus Holmlund. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I.* Cambridge University Press chapter Performance Information and Voting Behavior in Burkina Faso's Municipal Elections: Separating the Effects of Information Content and Information Delivery, pp. 221–256.

Lindbeck, Assar, and Jörgen W. Weibull. 1987. "Balanced-Budget Redistribution as the Outcome of Political Competition." *Public Choice* 52 (3): 273–297.

Manski, Charles E. 2003. *Partial Identification of Probability Distributions.* New York: Springer.

Morris, G. Elliott. 2018. "2018 U.S. House Midterm Elections Forecast." Available at `https://www.thecrosstab.com/project/2018-midterms-forecast/`.

Ofosu, George Kwaku. 2019. "Do Fairer Elections Increase the Responsiveness of Politicians?" *American Political Science Review* First View: 1–17.

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.

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.

Slough, Tara, and Christopher J. Fariss. 2020. "Misgovernance and Human Rights: The Case of Illegal Detention without Intent." *American Journal of Political Science* Forthcoming.

Smith, Alastair, and Bruce Bueno de Mesquita. 2011. "Contingent Prize Allocation and Pivotal Voting." *British Journal of Political Science* 42: 371–392.

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.

Willis, Derek. 2014. "Professors' Research Project Stirs Political Outrage in Montana." *New York Times*, Available at: `https://www.nytimes.com/2014/10/29/upshot/professors-research-project-stirs-political-outrage-in-montana.html`.

Appendices

Note: These supplementary materials are not intended for publication with manuscript.

A Vote Aggregation

Consider the case of an n-candidate (or n-choice) election in which n_d registered voters choose from candidates $i \in \{1, 2, ..., k+1\}$ where abstention is denoted by $k+1$. Vote totals absent the intervention are denoted v_d^i where $\sum v_d^i = n_d$. Without loss of generality, assume that:

1. Candidates 1 and 2 are the *ex-ante* marginal candidates.
2. $v_d^1 > v_d^2$ such that candidate 1 would win the last/only seat contested in the absence of intervention.[1]

From the definition of the margin to pivotality, ψ_d, $\psi_d n_d \equiv v_d^1 - v_d^2$.

In response to an intervention, denote the *net* change in votes from party r to party s as Δ_{rs}^d where $r < s$. If $\Delta_{rs} > 0 (< 0)$, candidate r received more (less) votes from candidate s voters than candidate s received from candidate r voters. The post-intervention vote total for party i, $\tilde{v}_d^i$, can thus be calculated:

$$\tilde{v}_d^i = v_d^i + \sum_{r=i} \Delta_d^{rs} - \sum_{s=i} \Delta_d^{rs} \tag{A1}$$

The difference in votes between candidate 1 and candidate i can thus be written:

$$\tilde{v}_d^1 - \tilde{v}_d^i = v_d^1 - v_d^i + \sum_{r=1} \Delta_d^{rs} - \left(\sum_{r=i} \Delta_d^{rs} - \sum_{s=i} \Delta_d^{rs} \right) \tag{A2}$$

If the intervention does not change the election result, candidate 1 must still win the last/only seat. This implies that $\tilde{v}_d^1 > \tilde{v}_d^i$ for all $i \in \{2, ..., n\}$.[2]

$$\tilde{v}_d^1 > \tilde{v}_d^i \Rightarrow v_d^1 - v_d^i > -2\Delta_d^{1i} - \sum_{r=1, s \neq i} \Delta_d^{rs} + \left(\sum_{r=i} \Delta_d^{rs} - \sum_{r \neq 1, s=i} \Delta_d^{rs} \right) \tag{A3}$$

Given an interference assumption, Definition 1, and the definition of Δ_d^{rs} imply that:

$$n_d MAEI_d \geq \sum |\Delta_d^{rs}| \tag{A4}$$

[1] In a PR election, it may be useful to think of v_d^i as a quotient or remainder on the last seat allocated. The logic follows equivalently.

[2] I assume that there is no minimal participation rule. Thus abstention (option $n+1$) therefore cannot "win," though this does not change the result.

Equation A4 further implies that $2n_d MAEI_d \geq 2 \sum |\Delta_d^{rs}|$. It therefore follows that:

$$2n_d MAEI_d \geq 2|\Delta_d^{i1}| + \sum_{r=1, s \neq i} |\Delta_d^{rs}| + \sum_{r=i} |\Delta_d^{rs}| + \sum_{r \neq 1, s=i} |\Delta_d^{rs}| \tag{A5}$$

Equations A3 and A5 imply that if $v_d^1 - v_d^2 > 2n_d MAEI_d$, it must be the case that $\tilde{v}_d^1 - \tilde{v}_d^2 > 0$. Substituting $v_d^1 - v_d^2 = \psi_d n_d$, if $\psi_d > 2MAEI_d$, then the experimental intervention could not change who wins office.

B Existing Experiments

I focus on published experiments on the provision of incumbent performance information to voters before elections, adapting the list of studies from Enríquez et al. (2019). Note that all calculations are back-of-the-envelope. I cannot estimate $E[a_{jc}(0)]$ in the case of cluster-randomized experiments. For this reason, I show the full range of $MAEI_d$ over the possible domain of $E[a_{jc}(0)] \in [0.5, 1]$.

Table A1 describes studies in the framework described in this paper.

Article	Country	Mapping to Framework	Calculation Details	$MAEI_d$ Est.	$\lvert D\rvert$
Adida et al. (2017)	Benin	d : Commune* c: Village (or urban quarters) j: Individual	Treatment (five variants) was assigned to 195 of 1498 villages. The saturation (density) of treatment in a village varies by treatment arm (below 100% in all villages) and village population is unclear without data on cluster size. Because of the noted distinction in the *de-jure* vs. *de-facto* characterization of parliamentary electoral districts, more information needed to clarify n_d.	–	30
Arias et al. (2019)	Mexico	d : Municipality c : Precinct j : Individual	At most $\frac{1}{3}$ of precincts per municipality were sampled, albeit non-randomly. Treatment was assigned to 200 households in each of 400 precincts (T1-T4). Precincts contain a maximum of 1,750 registered voters. Non-random sampling of precincts prevents calculation of $MAEI_d$.	–	26
Banerjee et al. (2011)	India	d : State leg. district c : Polling station j : Individual	20 treated polling stations and average of 57.5 control polling stations per district. All households were treated.	$[0.129, 0.258]$	10
Bhandari, Larreguy, and Marshall (2020)	Senegal	d: Department c: Village d: Individual	9 individuals were sampled per village. 450 villages were (non-randomly) sampled from the 859 villages in the 5 experimental departments. 375 villages received some treatment (non pure-control). Without further information on the distribution of villages (experimental and non-experimental) and population by district, the $MAEI_d$ cannot be calculated.	–	5
Boas, Hidalgo, and Melo (2019)	Brazil	d : Muicipality c : Individual j Individual	I assume $\frac{2}{3}$ of experimental sample was assigned to treatment (T1 or T2). The most over-sampled municipality had 416 voters in experimental sample and a population (not registered voters) of 45,503. If 70% of population were registered (mandatory in Brazil), upper bound (for any district) is given by $\frac{2}{3}\frac{416}{.7\times45{,}503}$.	0.008	47

Article	Country	Mapping to Framework	Calculation Details	$MAEI_d$ Est.	$\|D\|$
Buntaine et al. (2018)	Uganda	d : District c : Individual j : Individual	Study includes 16,083 subjects (T or placebo) in 111 districts. The subjects per district and registered voters per district are not provided so $MAEI_d$ cannot be calculated.	–	111
Chong et al. (2015)	Mexico	d: Municipality c : Precinct j : $Individual$	450 of 2360 precincts were treated (selected randomly). No information is provided on saturation within precinct so I assume all voters were treated.	$[0.095, 0.191]$	12
Cruz et al. (2019)	Philippines	d : Municipality c : Village j: Individual	All households in 104 treatment villages (T1 or T2) across 7 municipalities were visited. Each municipality has "20-25 villages." I assume 25 villages/municipality and that the experimental villages were randomly sampled.	$[0.279, 0.594]$	7
Cruz, Keefer, and Labonne (2018)	Philippines	d : Municipality c : Village j: Individual	All households in 142 treatment villages in 12 municipalities were visited. The average number of villages/municipality not reported. I assume 25 villages/municipality per Cruz et al. (2019) (which is consistent with 284 villages in the experimental sample). Villages were randomly sampled from the municipality.	$[0.237, 0.473]$	12
de Figueiredo, Hidalgo, and Kasahara (2011)	Brazil	d : Municipality c : Precinct j : Individual	$\approx$ All households were visited with flyers in 200 treatment (T1 or T2) precincts of 1,759 precincts in the municipality. The precincts were selected randomly subject to a set of constraints.	$[0.057, 0.114]$	1
George, Gupta, and Neggers (2018)	India	d : Assembly constituency c : Village j : Individual	The intervention treated 500,000 voters (T1-T4) in 1,591 villages. Villages have $\approx 1,200$ registered voters, so saturation rate in treatment villages was averaged 26%. Non-random sampling of villages within constituencies prevents estimation of $MAEI_d$.	–	38

Article	Country	Mapping to Framework	Calculation Details	$MAEI_d$ Est.	$\|D\|$
Humphreys and Weinstein (2012)	Uganda	d : Parliamentary constituency c : Polling station j: Individual	2 polling stations in selected constituencies and all households visited with flyers. Number of polling stations/constituency not reported so $MAEI_d$ cannot be calculated. The total number of constituencies where experiment occurred (known to be <147) is not reported.	–	–
Lierl and Holmlund (2019)	Burkina Faso	d: Village* c: Individual j: Individual	12 individuals were assigned to treatment (T or placebo) per village. Information about village population (n_d) is not reported.	$\frac{12}{n_d}$	146
Sircar and Chauchard (2019)	India	d : Assembly Constituency c: Polling booth area j : Individual	16 polling booth areas per precinct assigned to treatment (T1 or T2) with $\frac{2}{3}$ of households in each polling booth area assigned to receive flyer. While selection of experimental polling booths is random, the total number of polling booths per constituency is not reported so $MAEI_d$ cannot be calculated	–	25

Table A1: Survey of experiments on information disclosure about incumbent performance. * indicates that there may be distinctions between the *de-jure* electoral system and the *de-facto* vote aggregation rule, indicating some uncertainty about how to determine the electoral district.

C Supporting Information for Empirical Illustration

C.1 Data and Data Sources

I simulate different research designs on electoral data from the state of Colorado. Because statewide data it is sufficient to simulate all but presidential elections (and the Electoral College renders states the first unit of aggregation in presidential elections), I randomly selected the state of Colorado. As such data comes from:

- Colorado:
  - Precinct-level electoral returns voter registration from Colorado Secretary of State `https://www.sos.state.co.us/pubs/elections/VoterRegNumbers/VoterRegNumbers` `html`
  - 2018 House of Representative seat predictions from The Crosstab `https://www.thecrosstab.com/project/2018-midterms-forecast/`

		Registerd Voters		Precincts	
District type	Year	Mean	Std. Dev.	Mean	Std. Dev.
State House	2018	55,472	9,489	43.55	15.67
US House	2018	505,812	61,654	404.43	89.46

Table A2: Summary statistics on State and US House districts in terms of registered voters and precincts. Note that past electoral data from 2012, 2014, and 2016 is also collected for use in prediction.

C.2 Mapping the Framework onto Data

To clarify how the data is used, I map the parameters expressed in the paper onto variables in the data/simulation in Table A3.

C.3 Prediction Method

While much has been invested in predicting the results of national elections (in some countries), much less effort has been invested in predicting lower-level (state- and local-level) elections and elections in developing countries. In particular, there is a general lack of public opinion polling in these races. I consider what is possible to ascertain through registration data and past electoral returns alone. I propose estimate the predictive distribution of each ψ_d, following the steps:

1. Estimate a model of the form: $y_i = f(\boldsymbol{\beta}\mathbf{X}_i)$, where $\mathbf{X}_i$ is a matrix of predictors. Note that the unit of analysis is the district.
2. Generate many draws from the joint distribution of $\boldsymbol{\beta}$. For each draw:
   (a) Estimate $\widehat{\psi_d}$ from the model (possibly by aggregating over precincts). Then calculate $\widehat{\epsilon} = \psi_d - \widehat{\psi_d}$, the residuals, denote the pdf of residuals by $f_{\widehat{\epsilon}}$.
   (b) Randomly sample $x \sim f_{\widehat{\epsilon}}$ and calculate $\widehat{\psi_d} + x$.
3. These estimates form the empirical distribution $f(\psi_d|\widehat{\boldsymbol{\theta}})$.

Variable	Mapping	Notes
j	Individual voter	
c	Simulation varies for: {Individual, precinct}	Implies n_c
d	Given by the electoral district for a context	Implies n_d
S_{10}	Set of treated voters. Implied by specification of c and assignment of treatment.	
S_{00}	Set of untreated voters. Implied by specification of c and assignment of treatment.	
$E[a_c(0)]$	Bound on possible change in turnout.	Predicted from available data or set to maximum (1) or minimum ($\frac{1}{2}$) possible values for all precincts.
ψ_d	Predicted margin of victory in district d.	Predicted from available data or third-party prediction algorithm (in US Congressional elections only).

Table A3: Mapping of parameters of the model onto variables in the data and simulation. I assume that, as in Case #1, no intervention would happen in the absence of the experiment, i.e. $|S_{11}| = 0$ and $|S_{01}| = 0$.