The Ethics of Electoral Experimentation: Design-Based Recommendations

Tara Slough*

August 23, 2019

Preliminary and Incomplete

Abstract

While experiments on elections represent an increasingly popular tool in the social sciences, the possibility that experimental interventions could affect who wins office remains a central ethical concern. I argue that researchers should design electoral experiments to minimize the likelihood of changing such outcomes. This paper develops a formal characterization of electoral experimental designs that generates an upper bound on aggregate electoral impact under different assumptions about interference. I present a decision rule for comparing this bound to predicted election outcomes by which researchers can decide whether an experimental design should be implemented. I show that 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 recommendations identify trade-offs between adhering to ethical commitments and the knowledge generated by some types of electoral experiments, which I demonstrate by simulation on real electoral data. In sum, this paper advances an argument that some ethical concerns with experiments should be addressed through careful research design.

*Ph.D. Candidate, Columbia University; Predoctoral Fellow, University of California, Berkeley; and Visiting Scholar, New York University, tls2145@columbia.edu. I thank Macartan Humphreys and John Marshall 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 an increasingly 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 elections 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 largely *outside* jurisdictions where they work or plausibly reside. As a discipline, we now work far beyond elections in Chicago, Ann Arbor, and New Haven. In this sense, researchers will not typically internalize the costs or benefits of having altered an election outcome, should their intervention do so. However, the subjects of their experiments will internalize these outcomes.[1] Second, the scale of electoral interventions, measured in terms of the number of treated voters, has increased precipitously since early experiments. In addition ot 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). However, it remains conspicuously absent from larger disciplinary discussions of experimental ethics (e.g., Desposato, 2016). Indeed, this consideration appears to be invoked informally, if at all, in most *ex-post* accounts of electoral interventions. This article proposes a formal, design-based approach to the *ex-ante* consideration of how experimental interventions

[1]This is a feature of all electoral interventions to be sure, yet we may be particularly attuned to such issues when the "distance" between experimenters and subjects increases.

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, Desposato (2018) and Teele (2019) raise questions about standards for consent in field experiments including those on elections and Beerbohm, Davis, and Kern (2017) argue that experimentation in elections may undermine political equality in general. Carlson (2019) notes that researchers must weigh these epistemic benefits against the potential for harm to subjects. While these points merit a lengthier discussion, the focus of this paper is to consider how experimenters can minimize the risk of changing election outcomes, thereby doing harm to some subjects.

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. Given the frequent omission of the relationship between the district and the experimental units (of assignment or outcome measurement), it is often difficult to estimate *ex-post* the saturation of an intervention in the relevant electorate.

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 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 ignores 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 this paper, 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.

This analysis identifies three principal 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 (3) avoid manipulating interventions in close or unpredictable contests. Yet, these design principals come 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. Further, these design principals may limit the types of interventions we attempt to study. Interventions that vary the saturation of an intervention by cluster to study phenomena like coordination are particularly risky. Finally, strategies (2) and (3) suggest a trade-off between minimizing electoral impact and external validity.

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 th 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*, I suggest that 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 and Mexico, simulating experimental designs to compare possible electoral impact. 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 in informing experimental research design choices. While some existing works adopt non-standard experimental designs as a function of ethical considerations (i.e., Slough and Fariss, 2019), 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 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 harms and benefits to subjects, 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 typically harm some subject (i.e., a candidate made to lose support) while accruing benefits to another subject (i.e., a candidate with increased support). Even if we were to restrict the subject designation to voters, so long as preferences vary across the electorate, the possibility of shifting support from one candidate to another ostensibly generates harm and benefit to different groups of electors. In this sense, we know that electoral interventions generate harms to some actors.

[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).

The aggregation of votes to determine outcomes introduces two normative considerations unique to the electoral context. First, it suggests that the set of individuals that realize the consequences of an intervention includes members of a district, which often surpasses the number of experimental "subjects." Second, it weakens arguments in favor of standard efforts to mitigate experimentally-induced harm to subjects. In other (heavy touch) field experiments that could generate harm to (some) 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." Yet, the differences in the non-experimental and experimental allocations of a treatment combined can lead to distinct distributional outcomes as a consequence of the aggregation of votes by district.

2.1 Experiments and their Counterfactuals

A focus on the consequences of experimental interventions in elections requires consideration of what would happen in the absence of the experiment. Field experiments are often justified on ethical grounds by researchers who claim that the intervention would have occurred in the absence of an experiment and that the random assignment of treatment allows us to gain knowledge (Teele, 2013). In electoral settings, such justifications appear less frequently than in the program evaluation context. As such, a specified intervention may or may not be conducted 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 density 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 processes through which electoral experiments may be designed focusing on relevant actors in the research design process. I consider three (potential) actors: a researcher, an NGO, and an interest group. Researchers conduct an experiments primarily to learn

Class	Actors R	Actors NGO or IG	Experiment	Counterfactual (absent experiment)	Example
1	✓		*Researcher designs, implements experimental intervention.* (Note: A "partner" NGO or IG 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 cited in Dunning et al. (2019)
2	✓	✓	*Researcher randomizes an NGO or IG-funded and implemented intervention.*	*NGO or IG 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. Note the actors are: R = researcher; NGO = NGO; and IG = interest group.

something about the world. NGOs and interest groups are entities that participate in elections in some capacity, albeit with different objectives.[3] An NGO does not explicitly seek to change electoral outcomes for or against a specific candidate or position; in contrast, an interest group seeks to change such outcomes.[4] I assume that NGOs or interest groups that operate in an election do so in accordance with local electoral laws, such that their actions in the absence of an experiment can be seen as consistent with regular aspects of a campaign/election.

Table 1 identifies three classes of electoral experiments, delineated by the set of actors involved in their design and implementation. As is evident from comparison of the experiment and counterfactual columns, the experiment adds the researcher as an actor. Delineating the actors involved in an intervention proves instructive for understanding how the imposition of an experimental (randomized) allocation affects the distribution of a treatment.

First, consider the modal case (Case #1) of electoral experiments in which a researcher designs and implements an intervention that would otherwise not have occurred. A researcher values

[3]In most countries, "NGO" is a legal designation. The legality of NGO campaigns on behalf of a candidate or position varies accordingly. Here my distinction is based on an organization's objectives, not a legal designation. A registered NGO that campaigns for a candidate or position is an interest group under this definition.

[4]In some cases researchers collaborate directly with a candidate/campaign (e.g., Pons, 2018). For the purposes of this discussion, a campaign can be considered as analogous to an interest group.

learning; the most common quantitative operationalization of learning in experimental design is statistical power.[5] 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 in an experiment also increases the possible electoral impact of an intervention. 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 NGO or IG 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 expand into additional districts. 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 within a cluster are treated under an experimental allocation. For example, if a campaign targets its message at probable swing voters, under an experimental allocation, such voters must be probabilistically assigned to treatment. Even holding constant the number of treated voters, the set of treated voters may change when incorporating the experimental allocation of treatment. The relevant ethical concern in this case is the possible change in votes triggered by the experimental treatment allocation,

[5]If learning consists of Bayesian updating, a measure of difference between prior and posterior may be a better operationalization of learning. Unfortunately, we lack consensus on how to specify prior beliefs (or whose priors matter), making such an analysis more fraught. Given that learning consists of both a possible change in the mean and dispersion of beliefs, a design that maximizes power approximates learning as a reduction in posterior variance.

relative to the non-experimental allocation of the intervention.

2.2 The Ethical Objective

I assume that researchers' 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).

The approach advocated here considers two types of uncertainty that we have as researchers. First, we lack the omniscience to determine whether electoral outcome is normatively better than another for constituents or the grounds upon which such a determination could be made. As discussed above, 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 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 seats or offices. As such, this paper advocates the estimation of conservative bounds on the *ex-ante* possible shift in vote share. These bounds can be calculated analytically and compared to 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 (Beerbohm, Davis, and Kern, 2017). 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 outcome predictions 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) in a way 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' potential outcomes are affected 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 NGO or IG. 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. 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 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 counterfactuals described in Table 1. Our main potential outcome of interest, $\pi(E)$ is whether an individual voter is assigned to receive a treatment. Note that $\pi(e)$

is generally not independent of E. 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 strata, described in Table 2. I use the notation S_{11}, S_{10}, S_{01}, and S_{00} to denote the set of voters in each stratum. 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)	$\pi(E)$	$\pi(\neg E)$	Case 1	Case 2
S_{11}	1	1	$\lvert S_{11}\rvert = 0$	$\lvert S_{11}\rvert \geq 0$
S_{10}	1	0	$\lvert S_{10}\rvert > 0$	$\lvert S_{10}\rvert \geq 0$
S_{01}	0	1	$\lvert S_{01}\rvert = 0$	$\lvert S_{01}\rvert \geq 0$
S_{00}	0	0	$\lvert S_{00}\rvert > 0$	$\lvert S_{00}\rvert \geq 0$

Table 2: Principal strata. Each individual j in an electorate belongs to one exactly one stratum. The cases refer to those described in Table 1. The $\lvert \cdot \rvert$ notation refers to the cardinality of each set, or the number of voters in each stratum.

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 2, the relevant strata are S_{10} – individuals exposed to the treatment when it is assigned experimentally – and S_{01} – individuals not exposed to the treatment *because* is assigned experimentally. The proportion of the electorate in a district exposed to an intervention due to the experiment, heretofore the *experimental saturation*, $\mathcal{S}_d$ can thus be written:

$$\mathcal{S}_d = \frac{\sum_{c \in d} \sum_{j \in c} I[j \in \{S_{10} \cup S_{01}\}]}{n_d} \tag{1}$$

where $I[\cdot]$ is an indicator function. 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 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. The first – and weaker – assumption is simply the stable unit treatment value assumption (SUTVA), which is typically invoked to justify inference on 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:* $Y_j(z_{jc}) = Y_j(z_{jc}, \boldsymbol{z}_{j,\neg c})$

I add a second *within-cluster* non-interference assumption to the baseline calculation. Note that, in contrast to SUTVA, this is not a standard assumption justifying identification in cluster-randomized experiments. This assumption holds that, in the case that treatment is assigned to clusters of more than one voter, i.e. $|n_c| > 1$, 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.[6] 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 (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:* $Y_j(z_{jc}) = Y_j(z_{jc}, \boldsymbol{z}_{\neg j,c})$

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

3.3 Voter Response to the Treatment

Because the question at hand relates to whether an experimental intervention can change aggregate election outcomes, I focus on voting outcomes. To accommodate the range of interventions in the literature, I assume the potential outcomes framework as tractable and agnostic model of voting behavior for bounding outcomes. Specifically, given a treatment $z \in Z$, I assume that a turnout potential outcome $A_j(z) \in \{0, 1\}$ is defined for all j, z, where 0 corresponds to abstention and 1 correspond to a vote for some candidate.

While many electoral experiments seek to understand something about vote choice among voters beyond turnout, I focus on turnout for the construction of bounds for two reasons. First, turnout is conservative; any electoral outcome can be changed by a sufficiently large drop in vote turnout among some subset of voters. Second, a focus on turnout allows for abstraction from specific features of electoral systems.

I bound the plausible treatment effects on turnout 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 turnout outcome, one can bound the possible cluster-level treatment effects, among subjects whose treatment status is changed through the use of an experiment as: $ATE_c \in [-E[a_{jc}(0)|j \in \{S_{10} \cup S_{01}\}], 1 - E[Y_{jc}(0)|j \in \{S_{10} \cup S_{01}\}]]$. Note that these bounds are effectively "extreme value" bounds (Manski, 2003). From these bounds on ATE_c, I bound the maximum change in turnout (the larger bound in absolute value) at the the cluster level as:

$$\tau_c = \max\{E[a_{jc}(0)|j \in \{S_{10} \cup S_{01}\}], 1 - E[a_{jc}(0)|j \in \{S_{10} \cup S_{01}]\} \tag{2}$$

$$= \begin{cases} E[a_{jc}(0)|j \in \{S_{10} \cup S_{01}\}] & \text{if } E[Y_{jc}(0)|j \in \{S_{10} \cup S_{01}\}] \leq \frac{1}{2} \\ 1 - E[a_{jc}(0)|j \in \{S_{10} \cup S_{01}\}] & \text{if } E[Y_{jc}(0)|j \in \{S_{10} \cup S_{01}\}] > \frac{1}{2} \end{cases} \tag{3}$$

Several features of Equation 2 are of note. First, when treatments are assigned at the indivudual level such that $n_c = 1 \forall c$, $\tau_c = 1$. Second, in the case when clusters are the unit of assignment, τ_c is unknown *ex-ante* (and unrevealed in treated units *ex-post*). Researchers can conservatively

set $\tau_c = 1$ implying the largest possible electoral impact. Alternatively, τ_c can be predicted from pre-treatment data, such as past electoral outcomes or administrative records.

Note that the calculation of τ_c represents agnosticism about the direction of treatment effects. We could reduce τ_c if we were to impose more assumptions about the effects of an intervention (i.e., monotonicity). Indeed, our ability generate predictive theories about turnout decisions is notoriously weak (Bendor et al., 2011; Shapiro and Green, 1994).

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 $MAEI_d$. 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:*

$$MAEI_d = \frac{\sum_{c \in d} \tau_c \sum_{j \in c} I[j \in \{S_{10} \cup S_{01}\}]}{n_d} \tag{4}$$

Comment 1. *Comparative statics. The $MAEI_d$ is decreasing in the size of the electorate; increasing in the cluster-level experimental saturation; and increasing in the maximum changes in turnout.*

Comment 1 posits several immediate implications. Most obviously, an identically designed 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. However, ostensibly due to researchers' desire to work in low-information contexts,

much recent focus has been placed on legislative or local elections. This result suggests that this decision carries greater risks of altering turnout or vote share.

Second, Comment 1 suggests that higher saturation of (the experimentally-manipulated) treatment implies greater potential effects on vote share. This introduces a trade-off between statistical power and the degree to which an experiment could alter aggregate electoral outcomes. Treating more individuals increases the saturation of treatment, possibly moving more votes. Moreover, a move from a individually-randomized to a cluster-randomized experiment requires many clusters for adequate power to detect effects. To the extent that researchers treat large proportions of voters in clusters, the saturation of treatment increases substantially. One implication of lack of individual-level outcome data is that the possibility of experimental interventions altering electoral outcomes increases substantially.

4.2 Assessing the Consequences of Electoral Interventions

The final implication of Comment 1 with respect to τ_c demands 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 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 $\psi_d > MAEI_d$, then an experiment could not change the ultimate electoral outcome. In contrast, if $\psi_d < MAEI_d$, the experiment *could* affect the ultimate electoral outcome.

Unlike the other parameters of the design, τ_c (in cluster-randomized experiments) and ψ_d (in all

experiments) are not knowable in advance of an election, when researchers plan and implement an experiment. Imputing the maximum possible value of $\tau_c = 1$ allows for construction of the most conservative (largest) 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 $MAEI_d > \psi_d$ and *any* experiment could always 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, τ_c. 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 (posterior) 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 $\widehat{F}^{-1}(\underline{\psi_d}) = 0.05$, 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 $MAEI_d < \underline{\psi_d}$.

This decision rule rules out intervention in very close elections entirely. It permits experiments with a high experimental saturation of treatment only in "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 posible 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.

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 treated" (where present) or untreated voters in treated 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:

$$MAEI_d^w = \frac{\sum_{c \in d} \tau_c^w n_c I\left[\sum_{j \in c} I[j \in \{S_{10} \cup S_{01}\}] > 0\right]}{n_d} \tag{5}$$

Two elements change from $MAEI_d$ to $MAEI_d^w$. First, the "saturation" of the experiment grows to include all voters in a cluster in which any voter's assignment status is changed by an experiment. Second, the larger extreme value bound on turnout, τ_c^w now incorporates the (predicted)

untreated turnout level of the whole cluster. In the context of randomized saturation designs, the most common experimental design in which not all subjects in a treated cluster are treated, $E[\tau_c] = \tau_c^w$ given random sampling of the cluster population. This condition 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. Morever, 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.[7] In order to account for between-cluster interference, a violation of SUTVA, I introduce a parameters $\pi_c \in [0, 1]$ (for each c) to measure researchers' *ex-ante* beliefs about the threat of response 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 affected population in control clusters.

$$MAEI_d^{wb} = MAEI_d^w + \underbrace{\frac{\sum_{c \in d} \pi_c \tau_c^b n_c I\left[\sum_{j \in c} I[j \in \{S_{10} \cup S_{01}\}] = 0\right]}{n_d}}_{MAEI_d \text{ of inter-cluster to proportion } \pi_c \text{ voters}} \tag{6}$$

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, $MAEI_d^{wb} \geq MAEI_d^w$. This expression implies that we must make some possibly non-conservative assumptions about intercluster spillovers to design an experiment that is consistent with the proposed ethical guidelines. In other words, without circumscribing π_c in

[7] 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.

some way, we would never satisfy the decision rule proposed in this article in a contested election. (See Appendix XX for a formal statement and proof.) As such, a researcher would never run an electoral experiment if she anticipated between-cluster spillover effects that could reach all voters, even aside from 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 (possible) 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 (calculation based on report of 100,000 flyers in Willis, 2014). Efforts to measure campaign response such as (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.

If outside actors accurately target general equilibrium responses inside treatment clusters, the bound in Equation 5 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.

5.4 Concurrent Elections

Beyond standard concerns about wider exposure cross-sectional exposure to the treatment is the role of concurrent elections. Given 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. Yet, considerations of concurrent elections can lead to profound differences in assessments of the risk of electoral experiments.

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 published studies, suggesting high variability in the $MAEI_d$ to the extent that it can be calculated. 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 conduct a series of research design simulations 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 provide a survey as to how frequently the information

necessary to (begin to) aggregate votes is reported and what the 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 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 or a *maximum* $MAEI_d$ given accessible statistics. The *average* $MAEI_d$ is explicitly inappropriate for 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 $MAEI_d$ in six of 14 studies, varying τ_c from its minimum of 0.5 to its maximum of 1 for all c in cluster-randomized experiments. 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. Note however, 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.

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). Importantly, the purposive sampling is well-justified from a design perspective and the constituency size is not important 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. Yet, the variation in Figure 1 suggests that research designs vary substantially on this dimension and justify these considerations.

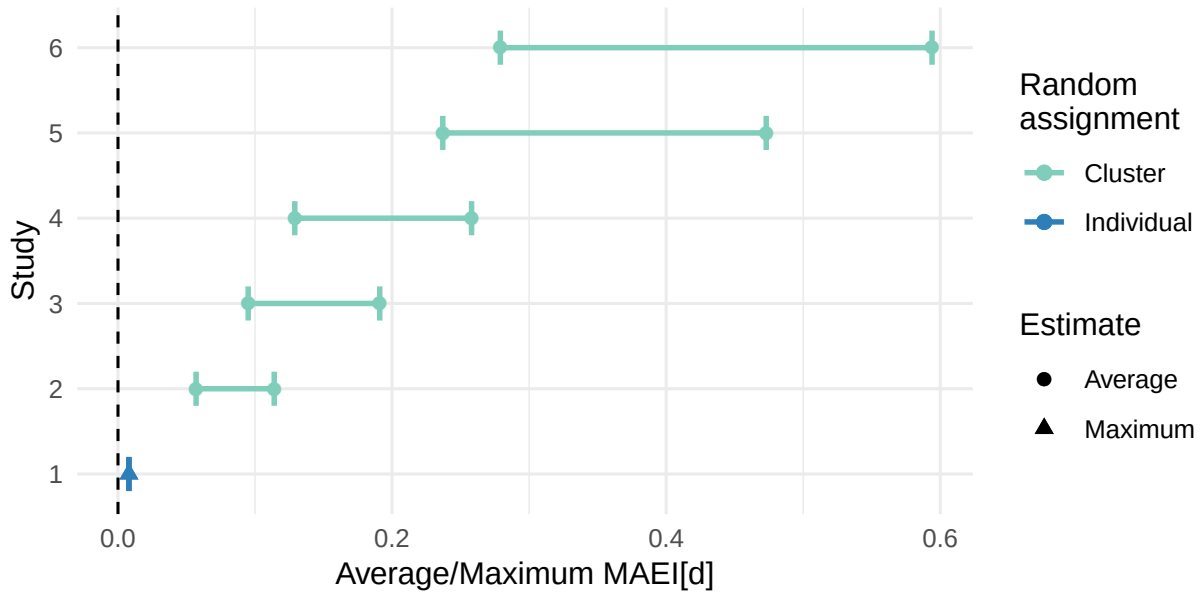

Figure 1: Estimated maximum or average $MAEI_d$ for six electoral experiments on electoral accountability. The interval estimates in the cluster-randomized experiments indicate the range of $MAEI_d$ estimates for any $\tau_c \in [0.5, 1] \forall c$. For discussion of these calculations, see Table A1.

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 and the Mexican state of Jalisco [TS note: Jalisco is not yet in this manuscript]. 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, the number of units assigned to treatment, and the allocation of treatments across districts.

In this version of the paper, I simulate a hypothetical experiments to be implemented in the Colorado 2018 election. I assume that these interventions would not occur absent the researcher. 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. Unfortunately, while much effort has been invested in prediction of 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

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

Table 3: Summary statistics on State and US House districts in terms of registered voters and precincts.

level offices or elections in developing countries. 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 predictive model on electoral data from 2012-2016 (three elections) and then predict outcomes for 2018.[8] I outline my prediction method 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% prediction 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 fall bound 0. More precise prediction algorithms may alleviate concerns in some of cases. On the other hand, elections are relatively "predictable" in the US given the comparative salience of partisanship in voting decisions and a vast amount of effort devoted to predicting and understanding US elections.

I consider several variants of research designs, each invoking SUTVA and, by design, satisfying Assumption 2.[9] To illustrate the various trade-offs, in Figure 3, I depict the maximum number of (1) individuals; and (2) clusters per precinct that could be assigned to treatment within the predictive bounds established above. In order to interpret these counts, Table 3 presents summary statistics.

Figure 3 suggests that higher proportions of individuals can be treated in less competitive dis-

[8]One concern is that this model does not incorporate time shocks absent polling data. This is one avenue for improvement in future iterations.

[9]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.

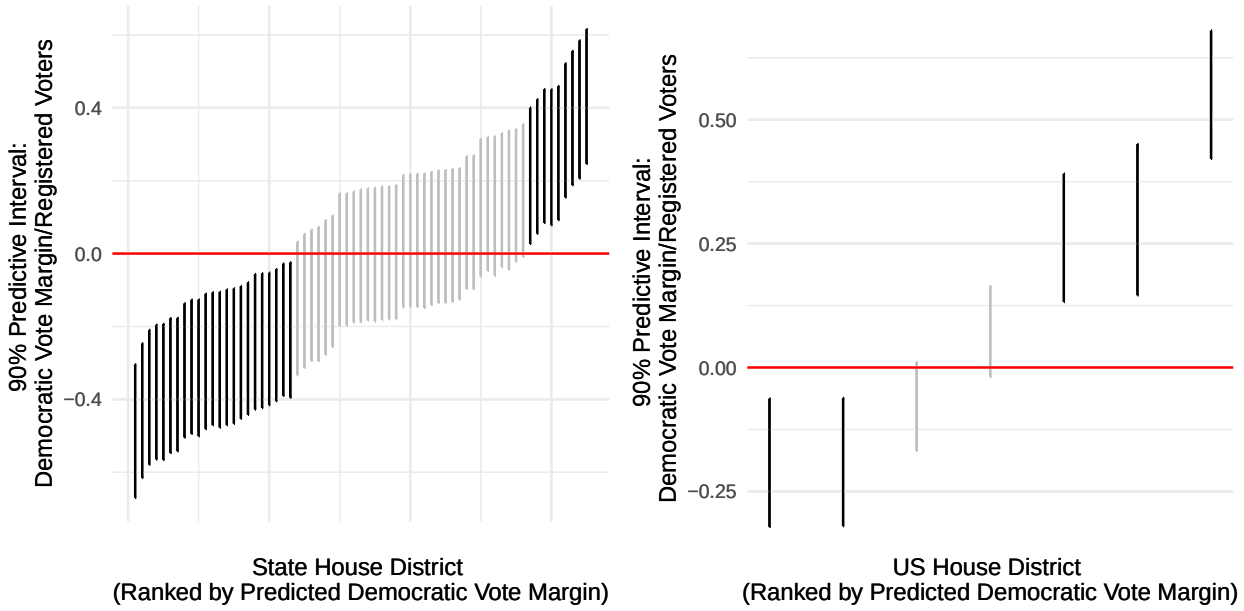

Figure 2: Predictive intervals for 65 State House seats and 7 US House seats in 2018. State House predictions are calculated following the method in Section C.3 while the US House predictions are "off the shelf" from Morris (2018). Grey lines represent grounds for declining to conduct an experiment in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

tricts. As described above, some districts (between the blue vertical lines) are removed from the experimental sample altogether given the possibility for close races. Striking, however, is how the counts in Figure 3 relate to the summary statistics in Table 3. Even in the most unequal contests, less than a third of population and precincts, respectively, should be assigned to treatment. In the case of individually-randomized experiments, experimenters are typically treating at rates far lower than those implied by this analysis. However, in the case of cluster-randomized experiments that treat at the precinct level, the saturation of treatment in electorates often far exceed the rates implied here. This observation mirrors the suggestive evidence on past studies in Figure A1.

[This section is preliminary and will ultimately include data from Jalisco and consider in more depth the implications of concurrent elections, a point omitted from the discussion.]

7 Implications for Research Design and Learning

The above discussion posits three main levers that researchers have to limit the possibility that their experimental interventions can change who wins elections:

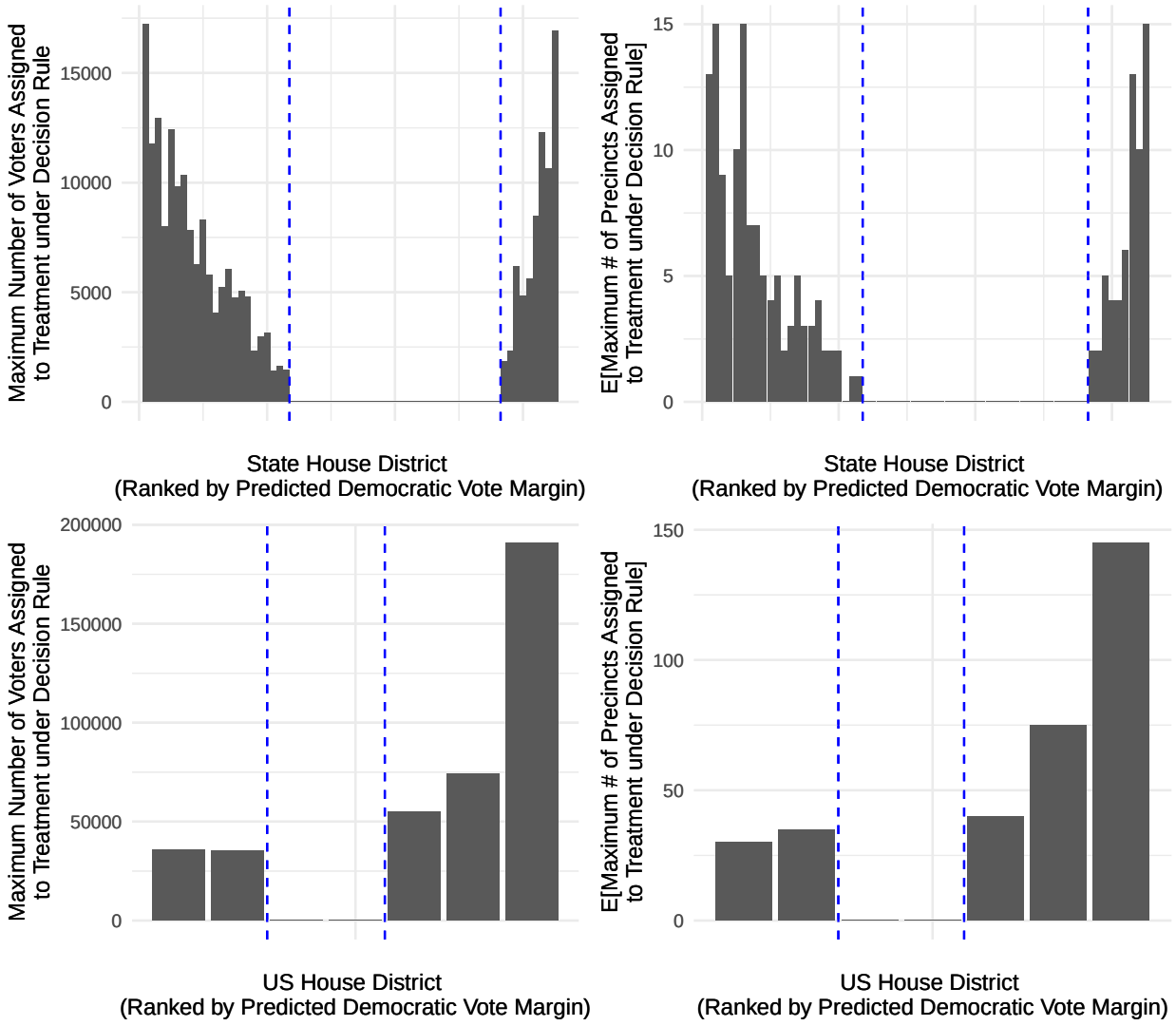

Figure 3: Maximum number of individuals (left) or precincts (right) that can be assigned to treatment under decision rule. For precincts, this reflects the expected number of precincts over multiple permutations of the assignment. The blue lines denote the districts excluded on the basis of predictions bounding 0.

1. Reduce the number of voters assigned to treatment.

2. Implement interventions in larger electoral districts.

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

These observations suggest trade-offs in the implementation of these research designs. First, consider the implications of #1 for statistical power. Constraints on power, at least in the number of observations, typically emerge from inability to treat enough individuals or clusters due to budgetary constraints. That number may be further constrained by concerns about minimizing electoral impact. This trade-off is particularly salient in experiments seeking to analyze aggregate electoral outcomes at the cluster (i.e. polling station or precinct) level. In such settings, researchers seek to ensure that a sufficient proportion of voters receives the treatment for treatment effects to be detectable. However, this requires the treatment of a much larger share of the electorate, particularly in small districts, which may be not satisfy the criteria specified in this article.

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 improvement in power for adding additional subjects is decreasing. Importantly, the possible electoral impact of an intervention increases linearly in the number of treated units, 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.

Collectively, recommendations #1-3 suggest that there are certain questions about elections that are challenging to address ethically in the context of experiments. For example, reducing the density of voters assigned to a treatment constrain the questions that we can address about saturation of interventions in an electorate; focusing on races with larger electorates may imply

fewer informational deficits on the part of voters, a central focus of recent experimental work; and avoiding close or unpredictable races reduces our ability to answers about how voters respond to pivotality concerns. Circumscribing the scope of what questions can be justifiably studied with experiments on elections does not necessarily weaken the enterprise. Experiments represent one of many tools that can be used to answer questions; like any other tool, their applicability is limited.

Finally, to the extent possible, researchers should implement interventions in "uncompetitive" and predictable races to the extent that we can predict these characteristics *ex-ante*. This suggestion comes from increasing the "margin to pivotality," or ψ_d. In such races, the ability of an experiment to change who wins office is lower. However, the study of voter or politician behavior in landslide races or "core" districts, may be quite distinct from behavior in more competitive constituencies. Consider, for example, longstanding discussions about the relationship between turnout and vote share. Do voters behave differently when they are more or less likely to be pivotal? In this sense, researchers face a trade-off in terms of the possible risks to election outcomes and the generalizability of insights about behavior.[10]

8 Conclusions

This paper argues that researchers, campaigns, and corporations should more carefully design electoral experiments in order to minimize the risk of changing electoral outcomes. Drawing upon a framework constructed upon nonparametric extreme value bounds, I suggest that researchers can best reduce ethical concerns through the choice of electoral district and allocation of treatment therein. I advocate against high saturation of treatment within districts and experimentation in close or highly unpredictable electoral contexts.

My argument posits a need for broader *ex-ante* consideration of ethics with respect to the context in which we experiment, here elections. Importantly, building ethical considerations into research design more broadly allows forces researchers to specify ethical concerns and mitigate the potential for harm. Because potential harms to subjects are realized when we field interven-

[10]While critiques of the lack of external validity of experiments are widespread, the idea that ethical considerations may lead to a less "representative" sample is new, to my knowledge.

tions, guidelines for the ethical design of experiments are useful before an experiment is fielded. Their utility decreases in the *ex-post* discussion of experiments as consequences have already been realized for subjects. As such, I recommend that the guidelines presented here be used to inform early-stage potential research projects.

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 context, similar considerations can be undertaken in most field experiments. As such, I advocate the incorporation of ethical considerations as a much more prominent guide to research design than is presently described.

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.

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`.

Bendor, Jonathan, Daniel Diermeier, David A. Siegel, and Michael M. Ting. 2011. *A Behavioral Theory of Elections*. Princeton, NJ: Princeton University Press.

Bhandari, Abhit, Horacio Larreguy, and John Marshall. 2019. "Able and Mostly Willing: An Empirical Anatomy of Information's Effect on Voter-Driven Accountability in Senegal." 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. 2019. "Field Experiments and Behavioral Theories: Science and Ethics." *PS Political Science and Politics* 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.

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.

Desposato, Scott, ed. 2016. *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*. New York, NY: Routledge.

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.

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, 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.

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.

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.

Shapiro, Ian, and Donald P. Green. 1994. *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. New Haven: Yale 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. 2019. "Misgovernance and Human Rights: The Case of Illegal Detention without Intent." Working paper available at `http://taraslough.com/assets/pdf/Haiti_paper.pdf`.

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

A 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 τ_c in the case of cluster-randomized experiments. For this reason, I show the full range of $MAEI_d$ over the possible domain of $\tau_c \in [0.5, 1]$.

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

B Statement and Proof of Interference Result

Proposition 1. *If between-cluster interference is not bounded, when $\pi_c = 1 \forall c \in d$, $MAEI_d^{bw} > \psi_d$ for any ψ_d and no experimental design can pass the decision rule.*

Proof: Combining Equations 5 and 6, $MAEI_d^{bw}$ can be written:

$$MAEI_d^{wb} = \frac{\sum_{c \in d} \tau_c^b n_c I\left[\sum_{j \in c} I[j \in \{S_{10} \cup S_{01}\}] = 1\right]}{n_d} + \frac{\sum_{c \in d} \pi_c \tau_c^b n_c I\left[\sum_{j \in c} I[j \in \{S_{10} \cup S_{01}\}] = 0\right]}{n_d} \tag{7}$$

If $\pi_c = 1 \forall c \in d$, Equation 7 can be rewritten:

$$MAEI_d^{wb} = \frac{\sum_{c \in d} \left(n_c \left(\tau_c^w I\left[\sum_{j \in c} I[j \in \{S_{10} \cup S_{01}\}] = 1\right] + \tau_c^b \left(1 - I\left[\sum_{j \in c} I[j \in \{S_{10} \cup S_{01}\}] = 1\right]\right)\right)\right)}{n_d} \tag{8}$$

By inspection of 8, $MAEI_d^{wb}$ represents the population-weighted average of τ_c^w and τ_c^b. This implies that $MAEI_d^{wb} \in [0.5, 1]$. Substituting $MAEI_d^{wb}$ into the definitition of τ_c (Equation 3), evaluate at the district level by defining c as equivalent to d, i.e. $n_c = n_d$. For clarity, I maintain the d subscript:

$$MAEI_d^{wb} = \max\{E[Y_{jd}(0)], 1 - E[Y_{jd}(0)]\}$$

This implies that the extreme value bound is $[-E[Y_{jd}(0)], 1 - E[Y_{jd}(0)]]$ which is equivalent to $[0, 1]$. As $\psi_d \in [0, 1]$, it must be the case that $MAEI_d^{wb} \geq \psi_d$ in any contested election if $\pi_c = 1 \forall c$. ■

C Supporting Information for Empirical Illustration

C.1 Data and Data Sources

I simulate different research designs on electoral data from the states of Colorado (U.S.) and the state of Jalisco, Mexico. Because both nations' elections are administered by states, voter registration data is managed by states. Because statewide data it is sufficient to simulate all but presidential

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) assigned to 195 of 1498 villages. Density of treatment in a village varies by treatment arm (below 100% in all) and village population is unclear without data on cluster size. Because of 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	Sampled at most $\frac{1}{3}$ precincts per municipality, albeit non-randomly. Treated 200 households in each of 400 precincts (T1-T4). Precincts had, at most, 1,750 random 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 (2019)	Senegal	d: Department c: Village d: Individual	9 individuals sampled per village. 450 villages are (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 $\frac{2}{3}\frac{416}{.7\times 45{,}503}$.	0.008	47
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

Article	Country	Mapping to Framework	Calculation Details	$MAEI_d$ Est.	$\lvert D \rvert$
Chong et al. (2015)	Mexico	d: Municipality c : Precinct j : $Individual$	450 of 2360 precincts were treated (selected randomly). No information 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 were visited in 104 treatment villages (T1 or T2) across 7 municipalities. 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 were visited in 142 treatment villages in 12 municipalities. 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 visited with flyers in 200 treatment (T1 or T2) precincts of 1,759 precincts 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	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
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.	–	–

Article	Country	Mapping to Framework	Calculation Details	$MAEI_d$ Est.	$\lvert D \rvert$
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.

	Election Year			
	2012	2015	2016	2018
Colorado				
Governor				✓
US Senate			✓	
US House			✓	✓
State Senate			✓ (half)	✓ (half)
State House			✓	✓
County (simulated races)			✓	✓
Jalisco				
Presidential	✓			✓
Governor	✓	✓		✓
State Legislature	✓	✓		✓
Mayoral	✓	✓		✓

Table A2: Electoral data used in simulations, by contest and year.

elections (and the Electoral College renders states the first unit of aggregation in presidential elections), I randomly selected the state of Colorado. Jalisco posts electoral data at the precinct level publically in an accessible format; it was not randomly selected among Mexican states. As such all 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/`
- Jalisco:
  - Precinct-level electoral data from `http://www.iepcjalisco.org.mx/resultados-elect`

The data I use in these simulations (a subset of all available data) is sumarized in Table A2.

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 the following method for estimating the predictive distribution of each ψ_d:

Variable	Mapping	Notes
j	Individual voter	
c	Simulation varies for: {Individual, precinct, district}	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.	
τ_c	Bound on possible change in turnout.	Predicted from available data, heuristic, or set to maximum possible value $\tau_c = 1$.
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 $|S_{01}| = 0$.

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 y_i may be aggregated at the district or precinct level.
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 $f(\psi_d|\widehat{\boldsymbol{\theta}})$.

C.4 Validation of Prediction Method

[To be completed.]