Political Analysis (2024), **32**, 384–400
doi:10.1017/pan.2023.44

PA

ARTICLE

Making a Difference: The Consequences of Electoral Experiments

Tara Slough

Department of Politics, New York University, New York, NY 10012, USA; Email: tara.slough@nyu.edu

(Received 31 July 2023; revised 15 October 2023; accepted 30 November 2023; published online 8 February 2024)

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. Researchers can mitigate the possibility of affecting aggregate outcomes by reducing the saturation of treatment or focusing experiments in districts and electoral systems where treated voters are less likely 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, this paper shows that the formalization of an ethical objective facilitates a closer mapping between ethical considerations and experimental design than is currently practiced.

Keywords: field experiments; elections; ethics

Edited by: Jeff Gill

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 have increased precipitously since the late 1990s. A central ethical concern in the study of electoral experiments is that by manipulating characteristics of campaigns, candidates, or voter information, researchers may also be changing aggregate election outcomes.

In contested elections, changing election outcomes through experimental interventions is apt to introduce downstream social harms to subjects and non-subjects alike, including candidates, their supporters, and some individuals that an election winner will ultimately govern. This concern is well documented in existing literature on experimental ethics. Teele (2013, p. 117) emphasizes the need for consideration of the "downstream and community-level risks" of field experiments. Phillips (2021, p. 281) emphasizes that: "the process-related downstream effects of these interventions can create winners and losers and harm individuals and groups. They can also harm entire communities." McDermott and Hatemi (2020), Gubler and Selway (2016), and Zimmerman (2016) elaborate particular concerns about the potential for disparate welfare impacts—including harms—of these downstream consequences across subjects, non-subjects, or communities in the context of elections. These consequences can be quite difficult to predict (Baele 2013; Carlson 2020).

© The Author(s), 2024. Published by Cambridge University Press on behalf of The Society for Political Methodology.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

Such arguments have led to ethical guidance that electoral experiments should be designed such that they are sufficiently unlikely to change aggregate election outcomes. Recently adopted American Political Science Association (APSA) "Principles and Guidance for Human Subjects Research" echo this concern, stating that interventions are of "minimal social risk if they are not done at a scale liable to alter electoral outcomes" (American Political Science Association 2020, p. 15). Similarly, Desposato (2016, p. 282) advocates "treading lightly."

To what extent do researchers adhere to this guidance when designing electoral experiments? Some authors report undertaking these considerations. For example, Dunning *et al.* (2019, p. 52) write that the authors of seven coordinated electoral experiments "elaborated research designs to ensure to the maximum extent possible that our studies would not affect aggregate election outcomes." However, the assembled pre-registered experiments from the American Economics Association and Experiments in Governance and Politics registries documented in Appendix A1 of the Supplementary Material tell a different story. Of the 129 experiments classified, just two discuss the possibility of changing aggregate outcomes among their written *ex ante* design considerations (Supplementary Table A1). To remedy this discrepancy between guidance and practice, this paper proposes a tool for the design of experiments that are unlikely to change aggregate election outcomes.

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

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

I propose a framework for bounding the maximum aggregate electoral impact of an electoral experiment *ex ante*. I focus on the design choices made by researchers, namely the selection of districts (races) in which to implement an intervention and the saturation of an intervention within these electorates. 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 (spillovers), 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 quantity compares to predicted electoral outcomes in a district, as formalized by a decision rule that can be implemented to determine whether to run an experimental intervention.

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) proportional representation (PR) contests. Importantly, the feasibility of these recommendations is conditioned by other institutional features of elections including district magnitude and concurrent elections, making some electoral interventions inherently more risky in some contexts than others. These design principles suggest trade-offs between ethical considerations and learning from electoral experiments. While some of these recommendations are intuitive and draw upon earlier contributions, the framework and decision rule provide the

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

first quantification of which designs treat "too many" voters while making explicit the assumptions underpinning these assessments.

This paper makes two central contributions. First, it provides a tool for researchers who are trying to avoid changing an electoral outcomes through experimental interventions. It helps identify the districts in which to intervene and determine appropriate saturation of treatment. These calculations can be implemented by an accompanying `R` package. Second, and more broadly, this paper makes a methodological contribution by showing how formalization of an ethical goal can bridge normative and experimental design considerations in order to address ethics concerns through experimental design.

2. Experiments and Their Counterfactuals

In order to consider the extent to which an experimental intervention might affect an electoral outcome, experimentalists must ask "what would have happened absent the experimental intervention?" The answer to this question generally depends on whether researchers are implementing their own intervention or randomizing an intervention that a partner would have implemented regardless. In the context of elections, these partners are typically political parties, NGOs, or candidates. The American Political Science Association (2020) and Hyde and Nickerson (2016) suggest that the ethical considerations with these partnerships depart from those of researcher-initiated and implemented interventions, advocating a lesser level of researcher responsibility.

The relevant consideration in cases of partnerships is how aggregate outcomes may be changed by *random assignment* of an intervention. Partners' pre-existing electoral goals likely guide how partners target interventions outside of an experiment. For example, an anti-corruption partner organization may prefer to target districts where corruption is worse or competitive districts where less corrupt candidates stand a better shot at winning. By randomly assigning the intervention, however, researchers may move the intervention away from the races in which it stands the best shot at achieving a partner's stated goal to reduce corruption. In this setting, the use of random assignment to assign a partner's well-intentioned and effective intervention may *reduce* welfare of subjects and non-subjects in the electorate relative to its non-experimental allocation.[1] Given these potential harms, I argue that in collaborations, researchers are responsible for how the random assignment of the intervention—in the service of research—changes the allocation of the treatment. In this sense, when collaborating with partners, researchers may be justified in studying interventions intended to change electoral outcomes, but these impacts should not be induced by the research component, specifically random assignment.

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

Ethical concerns about an experiment changing aggregate electoral outcomes should focus on the difference in treatment allocation between the experiment and its counterfactual. Because of the aggregation of votes at the district level (see Supplementary Table A4), subjects are not simply "interchangeable" when the allocation of treatment is changed. Therefore, within this framework, the guideline that "studies of interventions by third parties do not usually invoke [the principle of not impacting political outcomes]" put forth by American Political Science Association (2020, p. 14) is insufficient. Attention to the contrast between experimental and counterfactual allocation of an intervention arguably formalizes Hyde and Nickerson's (2016) concept of an intermediate level of scrutiny for experiments conducted with partners that is lower than the level of scrutiny afforded to experiments conducted without a partner.

[1] This raises a question of why a partner organization would participate in an experiment that might limit its efficacy. It may be the case that the partner: is willing to forego benefits today in the hope of learning for tomorrow; does not understand the implications of the reassignment of the intervention; or lacks information that would be needed to target the intervention better than randomly.

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

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

Case	Actors		Experiment	Counterfactual (absent experiment)	Examples
	Researcher	Partner			
1	✓		*Researcher designs and implements experimental intervention.* (Note: A partner may participate in or endorse the experiment, but the researcher causes the intervention to occur. Such interventions are often, but not necessarily, funded through the researcher.)	*No intervention occurs.*	Gerber and Green (2000); Metaketa-I experiments 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.*	Bond *et al.* (2012), Kendall, Nannicini, and Trebbi (2015), Pons (2018), López-Moctezuma *et al.* (2021)

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

2.1. The Ethical Objective

Following the American Political Science Association (2020) principles, I consider an ethical objective that aims to avoid changing who ultimately wins office. An experimentalist with this objective seeks to minimize the probability that their interventions change the *ex post* officeholders. This is clearly not the only relevant ethical consideration in the design of experiments, or even the only ethical goal with respect to aggregate outcomes.[2]

For example, the present objective assumes that the primary electoral consequences on policymaking or governance occur because candidate A wins office, not because candidate A won office with 60% instead of 51% of the vote. This objective abstracts from various effects of vote share (but not the winner) proposed in literatures on electoral autocracies (Simpser 2013) or distributive politics (Catalinac, de Mesquita, and Smith 2020; Lindbeck and Weibull 1987). However, the present framework provides the tools to rigorously develop alternate considerations. The calculation of aggregate electoral impact is not affected by the specific ethical objective. The decision rule, however, does depend on how this

[2]Appendix A4.4 of the Supplementary Material discusses concerns about impact on future election outcomes.

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

ethical objective is specified. Alternate objectives with respect to electoral outcomes can be formalized as different decision rules in cases where effects of changes in a winner's margin of victory pose particularly concerning implications for the welfare of voters, candidates, or district residents.

The framework starts from the observation that we can never 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. I advocate the estimation of conservative bounds on the *ex ante* possible change in vote share. These bounds are calculated analytically. I then develop a decision rule that compares these bounds to the predicted closeness of an election in order to minimize the risks of altering electoral outcomes. 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

Bounding aggregate electoral impact requires three considerations: design decisions made by researchers; researcher assumptions about which voters' behavior is affected by the intervention; and a minimal model of voter behavior that is sufficiently general to encompass many types of electoral experiments.

3.1. Research Design Decisions

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

Second, researchers define the clustering of subjects within a district. I assume that voters in district d, indexed by $j \in \{1,\ldots,n_d\}$ are partitioned into C_d exhaustive and mutually exclusive clusters. I index clusters by $c \in C_d$ and denote the number of voters in each cluster by n_c, such that $\sum_{c \in C_d} n_c = n_d$. There is always a cluster, even when treatments are individually assigned. Individual randomization can be accommodated by assuming $n_c = 1 \forall c$. Similarly, district-level clustering can be accommodated by assuming $n_c = n_d$.

Finally, researchers decide the allocation of the intervention 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. A subject's assignment to the intervention in state E is denoted $\pi(E)$. In the experiment, intervention assignment is random. Absent an experiment, the intervention could be assigned by any allocation mechanism. This notation allows for characterization of four principal strata, described in Table 2. By asserting the possibility of four (non-empty) strata, I allow for cases in which a researcher's partner would assign any proportion of the electorate (including all or none) to the intervention in the absence of the experiment. I use the notation S_{11}^{cd}, S_{10}^{cd}, S_{01}^{cd}, and S_{00}^{cd} to denote the set of voters belonging to each stratum in each cluster and district. The cases defined in Table 1 place assumptions on the relevant strata. Where the counterfactual is no intervention (Case #1), strata where $\pi(\neg e) = 1$ must be empty.

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

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

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

Table 2. 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.

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

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 exposed (resp. not exposed) to the intervention due to experimental assignment of treatment.

3.2. Researcher Assumptions about Interference between Voters

To construct bounds on aggregate electoral impact, researchers must make some assumptions about the set of voters whose voting behavior could be affected by an intervention. First, consider the stable unit treatment value assumption (SUTVA), which is invoked to justify identification of most standard causal estimands in experimental research. SUTVA holds 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 the treatment assignment of registered voter j in cluster c as $z_{jc} \in \{0,1\}$, the SUTVA for electoral outcome $Y_{jc}(z_{jc})$ is written in Assumption 1.

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

I add a second *within-cluster* non-interference assumption to the baseline model. Note that, in contrast to SUTVA, this assumption is not necessary for identification of standard causal estimands 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.[3] 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., Giné and Mansuri 2018; Ichino and Schündeln 2012; Sinclair, McConnell, and Green 2012), so I examine the implications of relaxing this assumption in Section 5.

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

3.3. Voters' Response to the Treatment

In order to ascertain whether an experimental intervention could change aggregate election outcomes, consider voting outcomes. Given treatment assignment z_{jc}, I assume that vote choice potential outcome $A_{jc}(z_{jc}) \in \{0,1\}$ is defined for all j and 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.).

[3]This assumption holds trivially in individually randomized experiments when $|n_c| = 1$ or when all registered voters in a cluster are assigned to treatment. It is relevant when only a portion of a cluster is assigned to the intervention.

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

Under Assumptions 1 and 2, 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, that is, 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. These possible ITEs serve as the basis for construction of extreme value bounds (EVBs) (Manski 2003).

To construct EVBs—and thus to calculate aggregate electoral impact—the expectation of untreated potential outcome of vote choice $E[a_c(0)]$ plays an important role. When treatment is cluster-assigned and $n_c > 0$, $E[a_c(0)]$ depends on which voters are assigned to the intervention. Random assignment of voters within a cluster ensures that $E[a_c(0)]$ is equivalent at varying levels of experimental saturation in treated clusters. This assumption can be relaxed when it is inappropriate, but the bound on aggregate electoral impact may increase.

EVBs can be very wide. Researchers may instead be tempted to use estimates from existing studies or less conservative bounding approaches. These approaches can be very misleading. Existing estimates are generally some form of average causal effect (i.e., an *ATE*). The ethical concern is not whether an experimental intervention changes outcomes on average. If *ATE* estimates that are small in magnitude mask heterogeneity, bounds based upon existing estimates will be non-conservative. Moreover, the outcome measures on vote share for a specific party or the incumbent (party) do not uniformly correspond to the relevant *ex ante* marginal winning (or losing) candidates, which means that they do not measure the effect that directly corresponds to considerations of aggregate electoral impact. Relative to other bounding approaches, I adopt EVB to avoid incorrect assumptions about the plausible effects of an intervention. This means that these bounds do not assume that an intervention will produce the hypothesized directional effect.

4. Bounding Effects on Electoral Behavior

4.1. Bounding Electoral Impact

Given these design choices, 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 quantity the *maximum aggregate electoral impact* in a district, the $MAEI_d$. Under Assumptions 1 and 2, this quantity is defined as follows:

Definition 1. Maximum Aggregate Electoral Impact: The ex ante maximum aggregate electoral impact (MAEI) in district d is given by

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

Consider the properties of $MAEI_d$ with respect to untreated levels of support for the winning candidate. Note that $E[a_c(0)] \in [0,1]$ for all $c \in C_d$. 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. This clarifies that researchers can generate a conservative measure of $MAEI_d$ simply by determining size of each stratum (in Table 2) in each cluster and district, without any additional electoral data or predicted outcomes. This means that researchers need only the number of registered voters in each district and cluster alongside their experimental design to calculate this quantity. 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

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

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. An otherwise identical experiment clearly has less possibility of moving aggregate vote share or turnout in a large district relative to a small district. Researchers' desire to work in low-information or low-turnout 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 aggregate impact, holding constant $E[a_c(0)]$. This suggests a possible trade-off between statistical power and the degree to which an experiment could alter aggregate electoral outcomes.

4.2. Assessing the Consequences of Electoral Interventions

The implications of $E[a_c(0)]$ for $MAEI_d$ prompt a discussion of the ability of electoral experiments to change electoral outcomes, that is, who wins. Analyses of electoral experiments typically focus on turnout or vote share, not the probability of victory (or seats won in a proportional representation system). The mapping of votes to an office or (discrete) seats implies the existence of at least one threshold, which, if crossed, yields a different set of office holders. For example, in a two candidate race without abstention, there exists a threshold at 50% that determines the winner. It is useful to denote the "*ex ante* margin of victory," 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, ψ_d represents the margin of victory (as a share of registered voters). 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 $\psi_d > 2MAEI_d$, then an experiment could not change the ultimate electoral outcome. In contrast, if $\psi_d \leq 2MAEI_d$, the experiment *could* affect the ultimate electoral outcome. Appendix A3 of the Supplementary Material shows formally the derivation of this threshold for an n-candidate race. The intuition behind the result is straightforward: $n_d\psi_d$ gives the difference in the number of votes between the marginal winning and losing candidates. The minimum number of votes that could change the outcome is $\frac{n_d\psi_d}{2}$, 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" margin of victory, 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 ψ_d 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 $\psi_d, \widehat{f}(\psi_d) \sim f(\psi_d|\widehat{\boldsymbol{\theta}})$, 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.[4] I propose the construction of a threshold based on the predictive distribution of ψ_d. Specifically, I suggest that

[4]I make no statement as to whether an experiment that passes the decision rule is ethical, since there are many other important ethical considerations for researchers to consider in addition to aggregate impact.

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

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 $2MAEI_d < \underline{\psi_d}$. While the possibility of changing 5% of election outcomes may seem non-conservative, recall that $MAEI_d$ captures the maximum possible effect of an intervention. It is highly unlikely that interventions achieve this maximum effect, reducing substantially the probability that an electoral experiment changes who wins an election.

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. Basing the decision rule on predictive distribution of ψ_d, as opposed to a point prediction, penalizes uncertainty over the possible distribution of electoral outcomes. Globally, the amount of resources and effort expended on predicting different elections and the quality of such predictions vary substantially. For this reason, I do not prescribe a predictive model. In some places, researchers may be able to access off-the-shelf predictions as in one case of the simulation below. In others, researchers would need to generate their own predictions using the data at hand. One implication of this variation in available information is that we are better able to make precise predictions in some electoral contexts than others. Where elections are highly unpredictable (due to lack of information, high electoral volatility, or limited election integrity), researchers concerned about aggregate impact should be more circumspect about experimental intervention. Use of this framework should alert researchers to the inherent risk of intervention in such contexts.

5. Allowing for Spillovers/Interference

Due to the use of extreme value bounds, decisions based on the $MAEI_d$ are conservative when Assumptions 1 and 2 are satisfied. By conservative, I mean that they will induce a researcher to err on the side of not conducting the experiment. Yet, when these assumptions do not hold, the same analysis might justify a non-conservative decision. For this reason, I examine the implications of relaxing these assumptions.

5.1. Within-Cluster Interference

One limitation of the previous analysis is that an intervention might only change the votes of those that are directly exposed within a cluster (Assumption 2). In this instance, clusters consist of multiple voters ($n_c > 1$), but not all voters in a treated cluster are treated or untreated due to the experiment. Yet, some "always assigned" (if 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 *et al.* 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

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

where $I[\cdot]$ denotes 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 voting behavior may be affected by the experimental intervention increases to include all voters in a cluster. 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 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

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

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 under SUTVA alone.

5.2. Between-Cluster Interference

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

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

The new term in the numerator of both expressions in (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 voting behavior of all registered voters in a district. In this case, (4) simplifies to

$$\max\left\{\frac{\sum_{c\in C_d}E[a_c(0)]n_c}{n_d},\frac{\sum_{c\in C_d}(1-E[a_c(0)])n_c}{n_d}\right\}. \tag{5}$$

However, it must always be the case that the *ex ante* margin of victory, $\psi_d \leq \frac{1}{n_d}\sum_{c\in C_d}E[a_c(0)]n_c$, 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 spillovers in some way, the decision rule is never satisfied in a contested election. Thus, a researcher should never run an electoral experiment if she anticipates between-cluster spillover effects that could reach all voters, even absent considerations of identification and inference.

6. Applications

I now apply the framework to existing experiments and conduct a simulation to show how one might use these tools to design an experiment.

6.1. Relation to Existing Electoral Experiments

I examine the application of this framework to existing experiments in two ways. In Appendix A5 of the Supplementary Material, I use replication datasets, administrative, and archival data to apply the framework to four published experiments that comprise mobilization, information, and persuasion interventions: Boas, Hidalgo, and Melo (2019), Gerber and Green (2000), Bond *et al.* (2012), and López-Moctezuma *et al.* (2021). And below, I focus on back-of-the-envelope calculation of the $MAEI_d$ on 14 studies of information and accountability that are classified by Enríquez *et al.* (2019). To compare these experiments, I use information reported in papers and appendices without consulting replication or

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

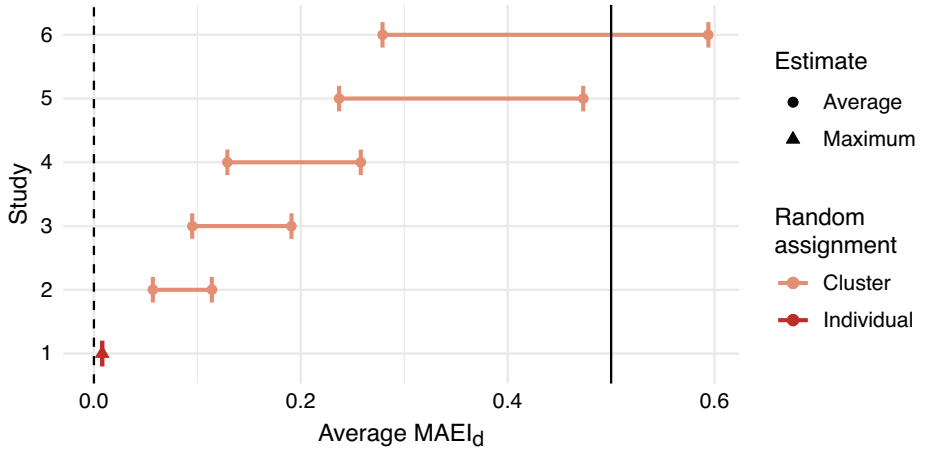


Figure 1. Estimated average $MAEI_d$ for six electoral experiments on electoral accountability. The interval estimates in the cluster-randomized experiments indicate the range of average $MAEI_d$'s for any $E[a_c(0)] \in [0.5, 1]$.

ancillary data. These calculations provide a survey of whether the information necessary to consider aggregate impact is reported, and what barriers to these calculations present. I report the studies, their relationship to the proposed framework, and my calculations in Supplementary Table A8. 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 calculate the $averageMAEI_d$ 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 serves as a measure of the variation in possible electoral impact. I am only able to estimate the $MAEI_d$ in 6 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 degree to which existing information experiments could have moved electoral outcomes varies widely. These back-of-the-envelope calculations, in isolation, cannot assess whether an intervention was consistent with the decision rule advocated here due to lack of information on the predicted margin of victory. However, any average $MAEI_d > 0.5$ cannot pass the decision rule (in at least one district) because $2MAEI_d > 1$. Supplementary Table A8 suggests that existing cluster-assigned information treatments tend to have high saturation 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 (four studies) or because there is insufficient information about constituency size, n_d (four studies). The takeaway from this survey of is simply that considerations of aggregate electoral impact require analyses that are not (yet) standard practice. The variation in Figure 1 further suggests that research designs vary substantially on this dimension and justify the considerations I forward.

6.2. Implementing the Framework

Simulating the design of experiments under the decision rule allows for an application of the full framework. The simulation below uses electoral data from the U.S. state of Colorado. It relies upon real voter registration data, precinct-to-district mappings, and election predictions. Because U.S. elections are administered at the state level, the simulations are greatly simplified by focusing on a single state in one election: the 2018 midterms. All races in 2018 were at the state level or below. In the simulations, I

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

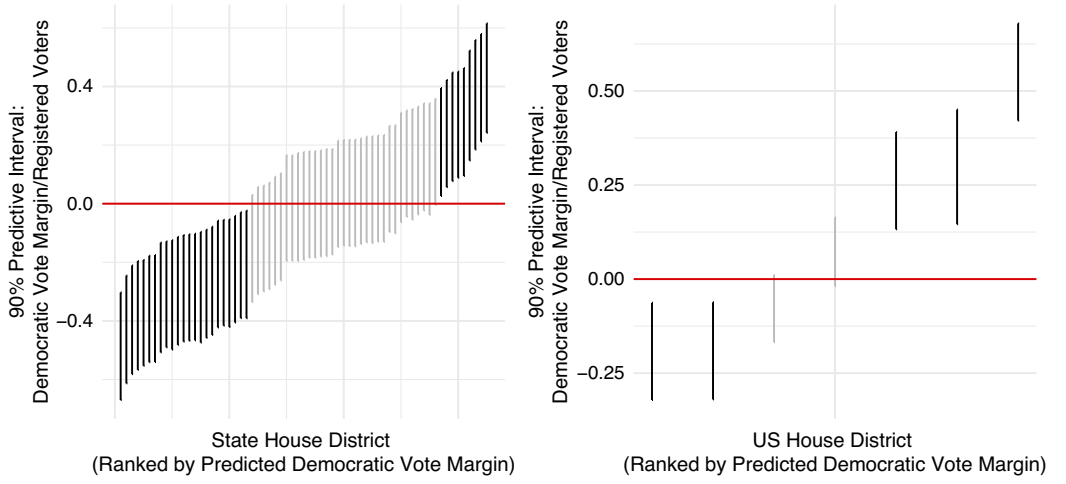


Figure 2. Predictive intervals for 65 State House seats and 7 U.S. House seats. Gray intervals represent grounds for declining to conduct an experiment in a district under the decision rule proposed here.

assume that an experimental intervention would not occur absent the researcher. I show that the method can be used with off-the-shelf predictions or with researcher-generated predictions. To illustrate the use of off-the-shelf predictions, I use the 2018 U.S. House forecast by Morris (2018). I generate my admittedly naïve predictions for Colorado State House seats in 2018 using (limited) available data, namely partisan voter registration data and lagged voting outcomes. I fit a basic predictive model on electoral data from 2012 to 2016 (three previous election cycles) and then predict outcomes for 2018 (see Appendix A7.3 of the Supplementary Material for details).

Examining only the predictive intervals, Figure 2 depicts the 90% predictive intervals for Colorado's 65 State House and 7 U.S. House seats in 2018. The 90% predictive intervals provide a useful visualization because when they bound 0 (the gray intervals), no experiment can pass the decision rule proposed in this paper. In sum, the predictive intervals for 33/65 State House races and 2/7 U.S. House races bound 0.

I consider two research designs, each invoking SUTVA and, by design, satisfying Assumption 2.[5] I first consider experiments that assign individual voters (not clusters) to treatment. I show calculations based on three types of 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 $E[a_c(0)] = 0$ and could represent the case in which all experimental subjects would vote in the same way absent treatment. This approximates a sample composed of only strong partisans. The intermediate case represented by "random sampling" predicts $E[a_c(0)]$ from 2016 district vote totals.

Figure 3 depicts the theoretical maximum number of individuals that could be assigned to an intervention in State House and U.S. 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, the more subjects can be assigned to treatment under the decision rule. Second, the type of experimental sample conditions the permissible treatment group size. However, going from worst to best case can double the number of subjects, as implied by (2). Third, comparing

[5]I assume all voters in cluster-randomized precincts are assigned to treatment if they belong to a treated cluster.

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

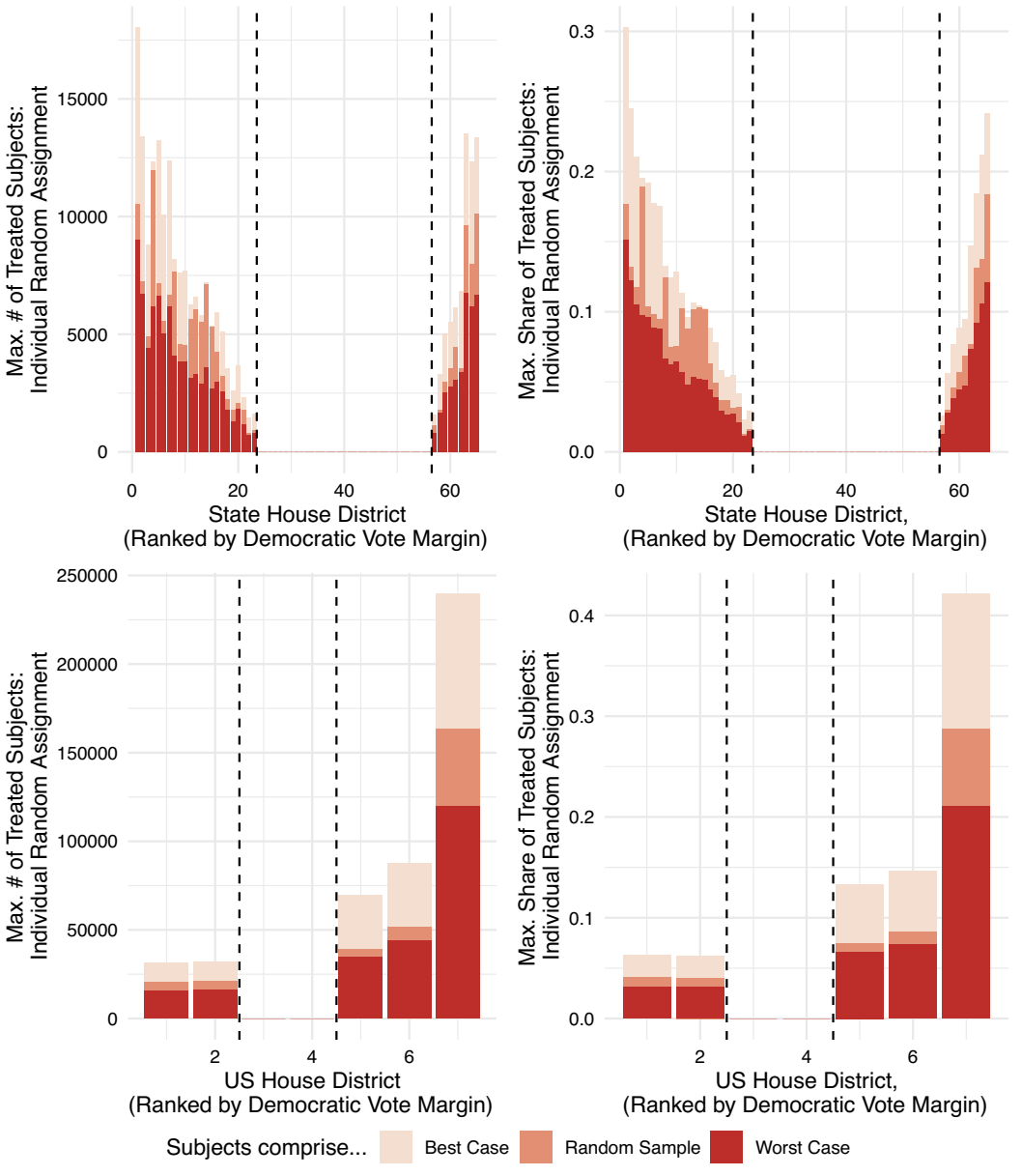


Figure 3. Maximum number of individuals (left) or proportion of registered voters (right) that can be assigned to treatment under decision rule.

the top and bottom plots in the left column, in larger districts, the maximum number of registered voters that could be assigned to treatment grows proportionately to district size (see Supplementary Table A9 for summary statistics). Finally, when describing the maximum number of treated subjects as a proportion of the electorate, only sparse treatments are permissible under the decision rule. Nevertheless, it implies that one could easily allocate an individually randomized treatment in a way such to power an experiment under the decision rule proposed by this paper. Supplementary Figure A11 reports the results of an analogous cluster-randomized treatment at the precinct level revealing that the number of "treatable" precincts is quite small, particularly in the case of State House races.

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

7. Implications for Research Design and Learning

The parameters used to characterize elections reflect features of both electoral systems, context, and data availability. I argue that best practices for electoral experiments are more likely to be tenable in some institutional contexts than others, as enumerated in Table 3.

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

1. Select treatments to improve the plausibility of assumptions of restricted interference.
2. Experiment in first-past-the-post (FPTP) races.
3. Implement interventions in larger electoral districts.
4. Avoid implementing experimental interventions in close or unpredictable races.
5. When intervening without a partner, reduce the number of subjects assigned to treatment in each district, d. When intervening with a partner, reduce the number of voters assigned to treatment or control because of the random assignment in each district, d.

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

These design strategies posit trade-offs in terms of learning from electoral experiments. I focus on implications for generalization and statistical power. Strategy #1 circumscribes the set of treatments that researchers develop and administer experimentally. In particular, this paper suggests that treatments

Table 3. Features of electoral systems and mapping to the framework.

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

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

that vary saturation of treatment assignment to study social dynamics or network effects of voting behavior are unlikely to pass the decision rule. Certainly, some social dynamics like coordination might be examined experimentally within small subsets of the electorate (i.e., within the household), but designs that examine these dynamics within large subsets of the electorate are less likely to pass the decision rule forwarded here.

Strategies #2 and #3 constrain the types of races in which intervention consistent with the ethical objective in this paper is feasible. These strategies rule out electoral experimentation in some countries or for some offices as a function of the electoral system or institutions. If voters, candidates, and parties behave differently under different electoral systems, there is not a theoretical expectation that results from one electoral experiment are informative about effects in a different context. To the extent that framework focuses on the ratio of treated voters to registered voters, the guidance to experiment in larger districts is similar to the guidance to reduce the number of voters assigned to treatment. However, it also speaks to the motivation of the interventions attempted. Many mobilization interventions focus on lower-turnout elections (i.e., primaries or local elections) and many information interventions focus on low-information elections (often local races). These approaches contrast with the guidance of Strategy #3.

Strategy #4—avoiding close or unpredictable races—posits concerns for generalization. This strategy excludes some polities with high volatility or minimal investment in election prediction. We may expect voters (or politicians) to act differently in places where a voter is more or less likely to be pivotal. If treatment effects vary in the characteristics used to target an experimental intervention, there exists a trade-off between these recommendations and our ability to assess the generalizability of insights about behavior.

Finally, Strategy #5 points to a familiar trade-off between statistical power and concerns about impacting aggregate electoral outcomes. I show that this trade-off is particularly salient in experiments seeking to analyze aggregate electoral outcomes at the cluster (i.e., polling station or precinct) level. At the same time, the framework provides novel guidance for the allocation of treatment across electoral districts. Specifically, it suggests that some power concerns may be reduced through higher levels of treatment saturation in districts where elections are highly predictable and lower levels of treatment saturation in districts where elections are predicted to be somewhat more competitive. The framework thereby provides new ways to improve statistical power given this trade-off.

Does the circumscription of electoral experiments to certain electoral contexts and treatments undermine the utility of electoral experiments as a tool? Here, an analogy to electoral regression discontinuity designs (RDDs) proves instructive. Electoral RDDs estimate some form of local average treatment effect at the threshold where elections are decided. The method is disproportionately used in low-level (i.e., municipal) FPTP contests. If these limitations on the application of electoral experiments are to be seen as damning to electoral experiments but not electoral RDDs, there seemingly exists an assessment that landslide races are less interesting—or of less political importance—than close contests. Theoretically, there are reasons why close contests may be reveal distinct strategic dynamics from predictable landslides. However, the ranking of these cases seems non-obvious. This paper simply advocates a more careful application of electoral experiments with recognition of their limitations, not a wholesale abandonment of the tool.

8. Conclusion

This paper shows that the formalization of an ethical objective can guide researchers to design research consistent with these standards. Similar formalizations of ethical objectives can guide the design of experiments beyond those in elections. An application of this approach to other experimental settings would consist of: (1) a clear mechanism linking the experimental intervention to relevant aggregate/societal outcomes; (2) a maximally agnostic model of how actors' responses to the intervention generate those outcomes; and (3) a set of assumptions restricting the set of actors that might respond to the treatment (the interference assumptions). Electoral experiments serve as an "easy" application

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

because elections offer a fixed, known mechanism for generating aggregate outcomes. On the other hand, in elections, the set of impacted actors (residents of a district) is often very large relative to the set of experimental subjects.

In sum, this paper suggests that by formalizing ethical goals and principles, researchers can better align experimental designs with these principles. Such methodological advances will allow researchers to continue to draw insights from the experimental study of elections while providing greater protections to the communities that we study.

Acknowledgments. I thank Jiawei Fu and Kevin Rubio for expert research assistance. I am grateful to Eric Arias, Graeme Blair, Alex Coppock, Sandy Gordon, Saad Gulzar, Macartan Humphreys, Kimuli Kasara, Dimitri Landa, Eddy Malesky, John Marshall, Lucy Martin, Kevin Munger, Gareth Nellis, Franklin Oduro, Melissa Schwartzberg, Dawn Teele, Lauren Young, two anonymous reviewers, APSA panel participants, and New York University graduate students for generous feedback. This paper was previously circulated under the title "The Ethics of Electoral Experiments: Design-Based Recommendations."

Funding Statement. This research was supported by a grant from the National Science Foundation (DGE-11-44155).

Competing Interests. The author has no competing interests.

Data Availability Statement. Replication code for this paper has been published in Code Ocean, a computational reproducibility platform that enables users to run the code, and can be viewed interactively at https://doi.org/10.24433/CO.7729631.v1 (Slough 2023a). A preservation copy of the same code and data can also be accessed via Dataverse at https://doi.org/10.7910/DVN/1UWD3M (Slough 2023b).

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2023.44.

References

American Political Science Association. 2020. "Principles and Guidance for Human Subjects Research." April 4. https://tinyurl.com/y5vm6cem.

Baele, S. J. 2013. "The Ethics of New Development Economics: Is the Experimental Approach to Development Economics Morally Wrong?" *Journal of Philosophical Economics* 7 (1): 2–42.

Boas, T., F. D. Hidalgo, and M. A. Melo. 2019. "Norms versus Action: Why Voters Fail to Sanction Malfeasance in Brazil." *American Journal of Political Science* 63 (2): 385–400.

Bond, R. M., et al. 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489: 295–298.

Carlson, E. 2020. "Field Experiments and Behavioral Theories: Science and Ethics." *PS: Political Science and Politics* 53 (1): 89–93.

Catalinac, A., B. B. de Mesquita, and A. Smith. 2020. "A Tournament Theory of Pork Barrel Politics: The Case of Japan." *Comparative Political Studies* 53: 1619–1655.

Desposato, S.. 2016. "Conclusion." In *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*, 267–289. New York: Taylor/Francis.

Dunning, T., G. Grossman, M. Humphreys, S. D. Hyde, C. McIntosh, and G. Nellis. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. New York: Cambridge University Press.

Enríquez, J. R., H. Larreguy, J. Marshall, and A. Simpser. 2019. "Information Saturation and Electoral Accountability: Experimental Evidence from Facebook in Mexico." Working Paper.

Gerber, A. S., and D. 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, X., and G. Mansuri. 2018. "Together We Will: Experimental Evidence on Female Voting Behavior in Pakistan." *American Economic Journal: Applied Economics* 10 (1): 207–235.

Gosnell, H. F. 1926. "An Experiment in the Stimulation of Voting." *American Political Science Review* 20 (4): 869–874.

Gubler, J. R., and J. S. Selway. 2016. "Considering the Political Consequences of Comparative Politics Experiments." In *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*, edited by S. Desposato. New York: Routledge.

Hyde, S. D., and D. W. Nickerson. 2016. "Conducting Research with NGOs: Relevant Counterfactuals from the Perspective of Subjects." In *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*, 198–216. New York: Taylor/Francis.

Ichino, N., and M. Schündeln. 2012. "Deterring or Displacing Electoral Irregularities? Spillover Effects of Observers in a Randomized Field Experiment in Ghana." *Journal of Politics* 84 (1): 292–307.

Kendall, C., T. Nannicini, and F. Trebbi. 2015. "How Do Voters Respond to Information? Information from a Randomized Campaign." *American Economic Review* 105 (1): 322–353.

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press

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

López-Moctezuma, G., L. Wantchekon, D. Rubenson, T. Fujiwara, and C. P. Lero. 2022. "Policy Deliberation and Voter Persuasion: Experimental Evidence from an Election in the Philippines." *American Journal of Political Science* 66: 59–74.

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

McDermott, R., and P. K. Hatemi. 2020. "Ethics in Field Experimentation: A Call to Establish New Standards to Protect the Public from Unwanted Manipulation and Real Harms." *Proceedings of the National Academy of Sciences* 117 (48): 30014–30021.

Morris, G. E. 2018. "2018 U.S. House Midterm Elections Forecast." November 6. https://www.thecrosstab.com/project/2018–midterms–f.

Phillips, T. 2021. "Ethics of Field Experiments." *Annual Review of Political Science* 24: 14.1–14.24.

Pons, V. 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, A. 2013. *Why Governments and Parties Manipulate Elections: Theory, Practice, and Implications*. New York: Cambridge University Press.

Sinclair, B., M. McConnell, and D. P. Green. 2012. "Detecting Spillover Effects: Design and Analysis of Multilevel Experiments." *American Journal of Political Science* 56: 1055–1069.

Slough, T. 2023a. "Replication Data for: Making a Difference: The Consequences of Electoral Experiments." Code Ocean. https://doi.org/10.24433/CO.77

Slough, T. 2023b. "Replication Data for: Making a Difference: The Consequences of Electoral Experiments." Harvard Dataverse, V1. https://doi.org/10.7910/DVN/1UWD

Teele, D. L. 2013. "Reflections on the Ethics of Field Experiments." In *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, edited by D. L. Teele, 67–80. New Haven: Yale University Press.

Zimmerman, B. 2016. "Information and Power: Ethical Considerations of Political Information Experiments." In *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*, edited by S. Desposato, 183–197. New York: Taylor/Francis.

Cite this article: Slough, T. (2024). Making a Difference: The Consequences of Electoral Experiments. *Political Analysis*, 384–400. https://doi.org/10.1017/pan.2023.44

https://doi.org/10.1017/pan.2023.44 Published online by Cambridge University Press