Making a Difference: The Consequences of Electoral Experiments

Supporting Information

Tara Slough*

December 20, 2023

Contents

*Assistant Professor, New York University. taraslough@nyu.edu

A1 Survey of Electoral Experiment Pre-Analysis Plans

A1.1 Justification

To survey the scope of electoral experiments, two forms of data collection seemed possible: a survey of published articles or pre-analysis plans. Sample selection is a concern in both cases, as depicted in Figure A1. The use of pre-analysis plans admits some bias toward recent studies since electoral experiments proliferated before the practice of pre-registering most experiments (c. 2013). This has the benefit of characterizing current practices without delays in publication. Further, it is useful to catalogue pre-analysis plans because we do not know the process via which the publication process selects on ethical concerns (or lack thereof). It may be that the most concerning studies are never written up or are rejected in the review process as a function of ethical concerns. By the same token, preferences for large-scale experiments or large samples may yield a sample of published studies that overrepresents the most concerning studies.

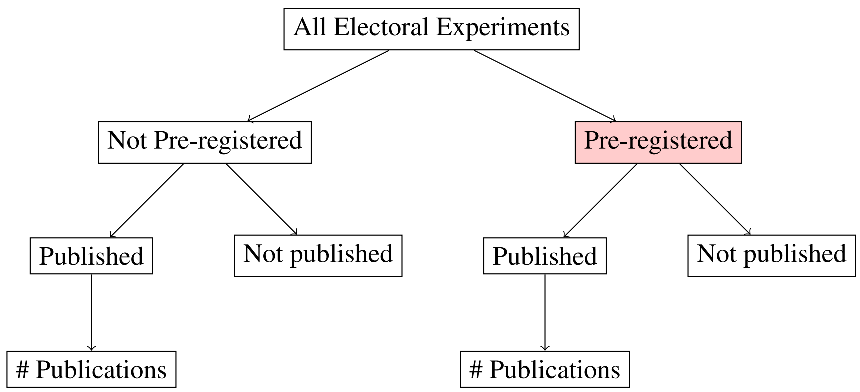


Figure A1: Selection processes. The current dataset emphasizes all pre-reregistered experiments in the AEA and EGAP databases, including all studies under the red node. Note that I only systematically observe the pre-registered experiments (and not their publication status).

I collect pre-analysis plans for electoral experiments that are registered in the Evidence in Governance and Politics (EGAP) and American Economic Association (AEA) registries. These registries are open to all and is costless for researchers (beyond the costs of preparing a pre-analysis plan). Reading registries and pre-analysis plans allows for systematic examination of how researchers addressed ethical questions when designing their experiments, when the method developed in this paper can be most productively implemented. While not all experiments are pre-registered in these registries, these documents provide some evidence about how researchers made design decisions and why. Most of the cited experiments in the paper (even some early studies) are included in this dataset of pre-analysis plans.

A1.2 Data Collection, Sample

I draw my sample of studies from the two largest social science registries: EGAP and AEA. At the time of database construction, EGAP had 1,434 registrations and AEA had 4,308. To narrow these registrations, I define an electoral experiment as one that:

1. Intervenes in a "real" election outside of a laboratory setting. This excludes a large number of survey experiments (typically vignettes) and lab-based studies of voting.

2. The researcher has some control over the random assignment of the intervention. This excludes natural experiments.

I consider downstream experiments based on the classification of the treatment. When authors look for downstream consequences of an ostensibly non-election related intervention (i.e., education interventions or a development pro-

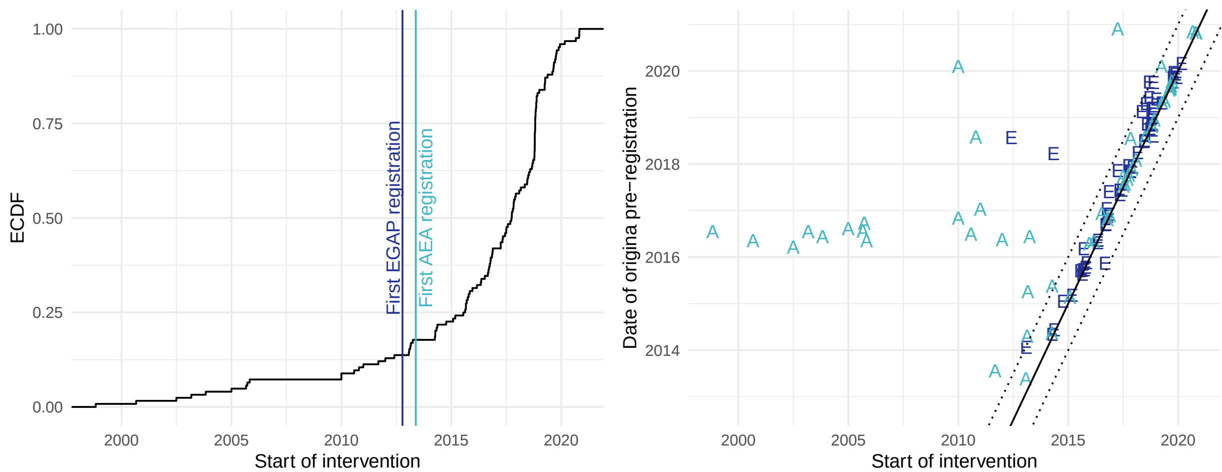


Figure A2: Distribution of the the start dates of electoral experiments in the pre-analysis plans dataset (left). Note that AEA encouraged the submission of old pre-analysis plans after the repository was created, accounting for many of the pre-2013 studies (right). In this plot "A" stands for AEA and "E" stands for EGAP

grams), I exclude them from the set of pre-analysis plans. If authors look at downstream outcomes of an election-related intervention I classify this as an elections experiment.

I describe the process through which the sample was sequentially narrowed in Table A1.

	EGAP		AEA	
Step	Criteria	n	Criteria	n
1	Complete database	1,434	Complete Database	4,308
2	Pre-2015: use of word "election" in database fields; post-2015 use of "election" keyword and "field experiment" methodology	209	Use of "electoral" keyword	181
3	Apply definition of electoral experiments (stated above)	77	Apply definition of electoral experiments (stated above)	57

Table A1: Selection of pre-analysis plans for inclusion from both databases.

For seven relevant studies, authors registered the studies in both registries. I remove the duplicates from the dataset for a total of 129 pre-registered electoral experiments. All duplicates are counted in the EGAP category from hereon.

Figure A2 summarizes the coverage of the pre-analysis plans. While researchers pre-register experiments in US elections at a higher rate than in any other country, the majority of experiments in the dataset are conducted outside the US. These AEA and EGAP registries became much more commonly used between 2015 and 2020 as pre-registration norms were popularized (Figure A2). Note that the AEA registry, in particular, contains some pre-analysis plans for experiments that were fielded *before* the creation of the registry, between 1998 and 2013. One further note is that a number of studies fielded in 2020 are currently embargoed and are thush not present in the dataset. These embargoes, in addition to the pandemic-related constraints on field research, may account for the slower pace of registrations in 2020.

Class	Description	Published Example	% of experiments in US	outside US
Mobilization	Get out the vote (GOTV) interventions. These interventions may use information to get out the vote, but they are aimed at mobilizing turnout, not changing votes or voter opinions of candidates/ballot items.	Gerber and Green (1999)	61.8%	23.0%
Information	Treatment is some form of information or access to information about candidates on the ballot or ballot items (as in referenda). Information may be prospective or retrospective.	Dunning et al. (2019), Bidwell et al. (2020)	14.5%	36.5%
Persuasion	Treatments intended to change votes or behavior in one specific direction. These experiments frequently randomize campaign strategy.	Wantchekon (2003), Kendall et al. (2015)	14.5%	18.9%
Election administration	Treatments intended to promote clean election administration or deter fraud.	Asunka et al. (2017)	–	10.8%
Voter registration	Treatments intended to induce eligible voters who are not registered to register to vote.	Shineman (2020)	5.5%	1.4%
Candidate-level	Treatments targeted at candidates or prospective candidates running for office.	Casey et al. (2021)	–	5.4%
Campaign finance	Treatments intended to shape financial contributions by voters	Boatright et al. (2006)	3.6%	–
Other	Other treatments deployed in advance of elections.	Jha and Shayo (2019)	–	4.1%

Table A2: Classification of experiments in the pre-analysis plan dataset.

A1.3 Intervention Classificaton

Table A2 elaborates the definition of each class and disaggregates the use of these interventions in the US and outside the US. Figure A3 depicts the distribution of intervention classses in the data.

A1.4 Ethics Coverage of Pre-Analysis Plans

Finally, I consider the coverage of ethical considerations in this subset of pre-analysis plans. In general, I consider three types of ethical considerations as they appear in pre-analysis plans or registry entries:

1. **IRB approval**: Does a registry entry or pre-analysis plan reports having IRB approval from at least one institution? Note that I do not assume that failure to include an IRB number does not necessarily mean that a project did not have IRB approval. For the AEA registry, this is a required entry, so the absence of an associated IRB protocol may be more indicative of a lack of approval in that dataset.

2. **Discussion of any additional ethical considerations**: Does the pre-analysis plan or registry discuss ethical concerns or risks beyond an IRB approval?

3. **Consideration of aggregate electoral impact**: Does the pre-analysis plan or registry discuss the possibility of changing election outcomes as a potential ethical concern?

For the EGAP registry, I reviewed pre-analysis plans for all registered experiments. For the AEA registry, I reviewed all public pre-analysis plans and registry entries (not all pre-analysis plans are public). For this reason, I disaggregate by repository before reporting the overall rate of inclusion of these items in Table A3.

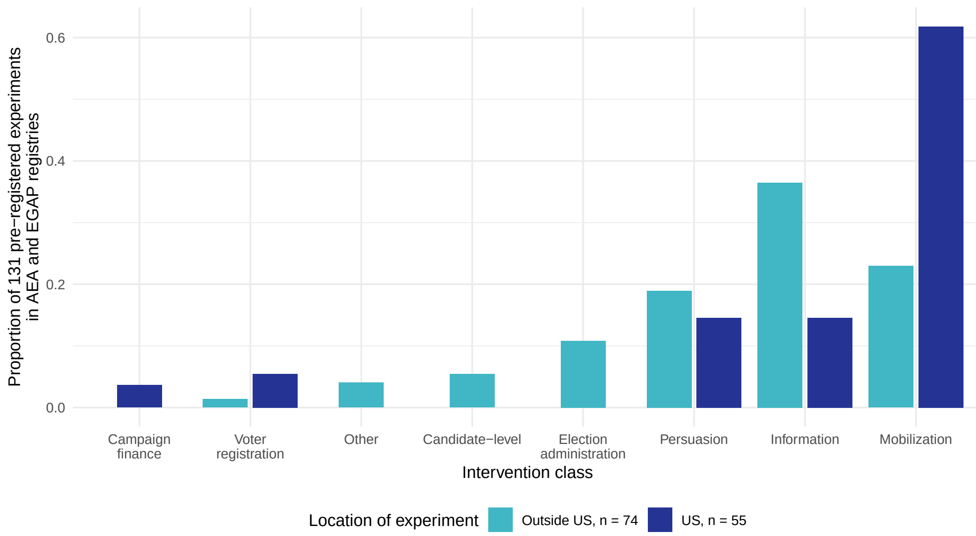


Figure A3: Types of experiments in pre-registered experiments in and outside the US. See Table A2 for clarification of the intervention classes.

Item	AEA registry ($n = 52$)	EGAP registry ($n = 77$)	Combined registries ($n = 129$)
IRB approval	73.1%	41.6%	54.6%
Additional ethical considerations	10.4%	0%	6.2%
Aggregate electoral impact	2.6%	0%	1.6%

Table A3: Ethical coverage of pre-analysis plans and registry entries.

A2 Counterfactual Allocations of Partners' Treatments

Table A4 provides an illustration of the three ways in which experimental allocations of a treatment may depart from implementing partners' allocation of the treatment outside of an experiment.

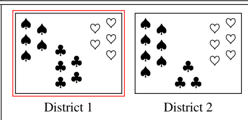
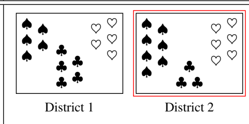
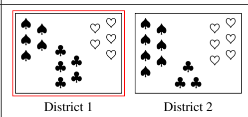
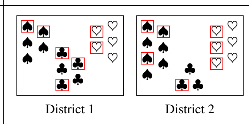
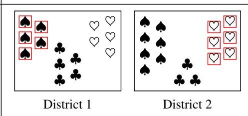
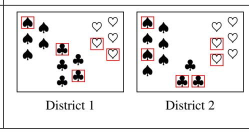
	Counterfactual allocation	Experimental allocation	Description
1			Intervention is implemented in **different districts** under experimental allocation.
2			Intervention is implemented at **different saturation** under experimental allocation.
3			Intervention is targeted to **different voters** under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

Table A4: Each symbol represents a voter. The red boxes indicate treatment assignment.

A3 Vote Aggregation: A Framework

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

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

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

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

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

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

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

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

$$\tilde{v}_d^1 > \tilde{v}_d^i \Rightarrow v_d^1 - v_d^i > -2\Delta_d^{1i} - \sum_{r=1, s\neq i} \Delta_d^{rs} + \left(\sum_{r=i} \Delta_d^{rs} - \sum_{r\neq 1, s=i} \Delta_d^{rs} \right) \tag{3}$$

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

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

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

$$n_d MAEI_d \geq \sum |\Delta_d^{rs}| \tag{4}$$

Equation (4) further implies that $2n_d MAEI_d \geq 2\sum |\Delta_d^{rs}|$. It therefore follows that:

$$2n_d MAEI_d \geq 2|\Delta_d^{i1}| + \sum_{r=1, s\neq i} |\Delta_d^{rs}| + \sum_{r=i} |\Delta_d^{rs}| + \sum_{r\neq 1, s=i} |\Delta_d^{rs}| \tag{5}$$

Equations (3) and (5) imply that if $v_d^1 - v_d^2 > 2n_d MAEI_d$, it must be the case that $\tilde{v}_d^1 - \tilde{v}_d^2 > 0$. Substituting $v_d^1 - v_d^2 = \psi_d n_d$, if $\psi_d > 2MAEI_d$, then the experimental intervention could not change who wins office.

A4 Discussion of the Ethical Objective

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

A4.1 Objections part I: The ethical objective is too permissive

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

Lack of consent: Like most field experiments, electoral experiments are generally conducted without the consent of subjects or non-subjects who may be affected by the intervention.[3] Teele (2013), Humphreys (2015), and McDermott and Hatemi (2020) note that the lack of consent in field experiments departs from standard requirements of informed consent in medical studies. Desposato (2018) shows that informed consent increases the proportion of American survey respondents and political scientists that find hypothetical experiments ethically acceptable. Objections to field experimentation on the basis of lack of informed consent extend far beyond electoral experiments. Humphreys (2015) and Teele (2019) chart a productive path forward concerning issues of consent. By conceptualizing consent more broadly, both authors provide new suggestions about how experimentalists might seek consent in new ways or provide additional protection to subjects and non-subjects in its absence. Further development of and debate about these alternatives is important to addressing issues of consent as a broader objection to field experimentation.

Electoral experiments violate self-determination: This paper makes a consequentialist argument: experiments on elections can generate social harm by changing who wins office. Baele (2013, p. 28) asserts that the primary deontological issue with electoral experiments is that they "influenc[e] political situations in other countries ... as it constitutes a breach in sovereignty if all the stakeholders do not agree in the process." Whitfield (2019, p. 7) argues more broadly for political research ethics that respect "self-determination of communities." As such, even if an experiment were to pass the decision rule I advance, violations of an community's (electorate's) right self-determination may justify broader restrictions on electoral experimentation. Two caveats are important. First, to my knowledge, this argument has not yet been developed in depth with specific reference to electoral experiments. Future work is needed to justify this highly restrictive stance on electoral experimentation. Second, it is unclear how these arguments apply to a researcher that implements an intervention in their own electoral district. Relatedly, Beerbohm et al. (2020) provides useful guidance on the moral status of treatments carried out by a researcher (presumably in their home district).

[3] While some researchers obtain the consent of (select) candidates or political parties, I am not aware of field-experimental electoral interventions that seek the consent of voters (subjects or non-subjects).

A4.2 Objections part II: The ethical objective is too strict

In sharp contrast, some readers may contend that the ethical guidance to not change election outcomes is too restrictive. Indeed, I have shown that there exist institutional and political contexts in which no experimental design is likely to pass the proposed decision rule. These obejections largely center two arguments. First, there may be benefits—either in the form of welfare or knowledge—that stem from electoral experiments. Second, the political science literature suggests that election outcomes may have many causes. Why, then, should we care so much about one (possible) cause: electoral experiments?[4]

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

The principle of beneficence holds that researchers should "maximize possible benefits and minimize possible harms" (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research., 1979, p. 5). Changing an electoral outcome will harm some individuals while benefiting others (Gubler and Selway, 2016; Zimmerman, 2016). For example, a candidate that loses *because of the intervention* and her supporters will generally be harmed by an intervention. By contrast, a candidate that wins *because of the intervention* and her supporters will generally benefit from the intervention. Researchers are often unable to anticipate or weigh the extent of the harms and benefits across an electorate (Carlson, 2020; Baele, 2013). When researchers lack foresight into the consequences of the intervention, they can minimize the potential for harm—while still gaining knowledge benefits—by designing the experiment to minimize aggregate electoral impact following the guidance in this article.

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

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

Other causes of election outcomes: There are arguably many causes of election outcomes. Why should we care so much about one potential cause—an electoral experiment—when other causes might be more influential (i.e., the economy) or more normatively concerning (i.e., sports game results)? The critical distinction between the experiment and "everything else" is that the experiment constitutes *research*, and is thus subject to research ethics.[5] In the process of doing research, researchers have a responsibility to protect subjects and their communities from possible social harms. Such responsibilities do not extend to non-research activities.

[4]Note that electoral experiments allow us to measure effects of causes, not causes of effects. Yet, we can still theorize about possible causes of these election outcomes.

[5]I follow the Belmont Report's definition of research as "an activity designed to test an hypothesis, permit conclusions to be drawn, and thereby to develop or contribute to generalizable knowledge" (p. 3).

A4.3 Minimizing aggregate electoral impact as a default

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

A4.4 Comment on future effects of an intervention

If researchers or consumers adopt the view that electoral experiments should not change who wins office, a further concern highlights the possibility that an intervention in election t could change electoral outcomes in election $t+1$. There are two potential mechanisms through which such effects could be realized. First, some actor in elections could learn from the experimental results and change their behavior in election $t+1$. Second, it is possible that an intervention has persistent effects on voter behavior in election $t+1$, through learning, habit formation, or incumbency effects.

The first mechanism, impacts through acquired knowledge, is addressed in American Political Science Association (2020): "These considerations and concerns for impact do not generally apply to the impact of knowledge generated by research activties" (p. 16). Many forms of research, far beyond human subjects research, in the social sciences could fall astray of this standard. Indeed, observational research on elections using only non-human subjects data could inform electoral behavior—and thereby electoral outcomes—through this channel. Rejection of this standard for impact surpasses considerations of human subjects research and is beyond the scope of discussions of beneficence.

The second mechanism, persistent treatment effects, relates to the framework developed in the paper. Under two assumptions: (1) holding fixed district boundaries and (2) maintaining the same assumptions about interference, the $MAEI_d$ should provide an upper bound on aggregate impact in election $t+1$. Therefore, this concern can be reduced to a concern that the election $t+1$ will be closer than election t. Where persistence of treatment effects is a concern, therefore, researchers should be extra circumspect of working in contexts with high electoral volatility.

A5 Applications of Decision Rule from Existing Literature

I walk through the application of the proposed framework and decision rule to four experiments documented in the existing literature. The goal is to illustrate four aspects of the current framework: individual vs. clustered random assignment and intervention without vs. with a partner (Case #1 vs. Case #2 of Table 1). The studies I employ are listed in Table A5.

	Without experiment:	
Assignment is:	No intervention occurs	Partner conducts intervention
Individually randomized	Boas et al. (2019)	Bond et al. (2012)
Cluster randomized	Gerber and Green (2000)	López-Moctezuma et al. (2021)

Table A5: Studies used for illustration of framework, decision rule.

In addition to calculation of the MAEI and (to the best of my ability) $\underline{\psi}_d$, I provide ex-post data on election results to highlight the relationship between *ex-ante* and *ex-post* voting.

A5.1 Application #1: Boas, Hidalgo, and Melo (2019)

Boas et al. (2019) conduct a field experiment in 46 municipalities in Pernambuco, Brazil prior to the 2016 election.

This experiment was one of seven experiments in EGAP's Metaketa-I on electoral accountability. It is notable because of the considerations of electoral impact undertaken in the pre-analysis plan.

1. **Electoral context**: This experiment provided information on municipal government performance to voters in 46 municipalities in the Brazilian state of Pernambuco prior to the 2016 municipal elections. Municipal elections include elections for mayor (FPTP races for all municipalities in the sample) and city council (open list PR races).

2. **Experimental overview**: The experiment consists of 3,200 voters, of whom 2,400 are assigned to two treatment conditions that reveal information about (1) municipal financial management and (2) education. The allocation of these 3,200 voters across 46 municipalities is depicted in A4. The experiment randomizes individual respondents to these treatment arms.

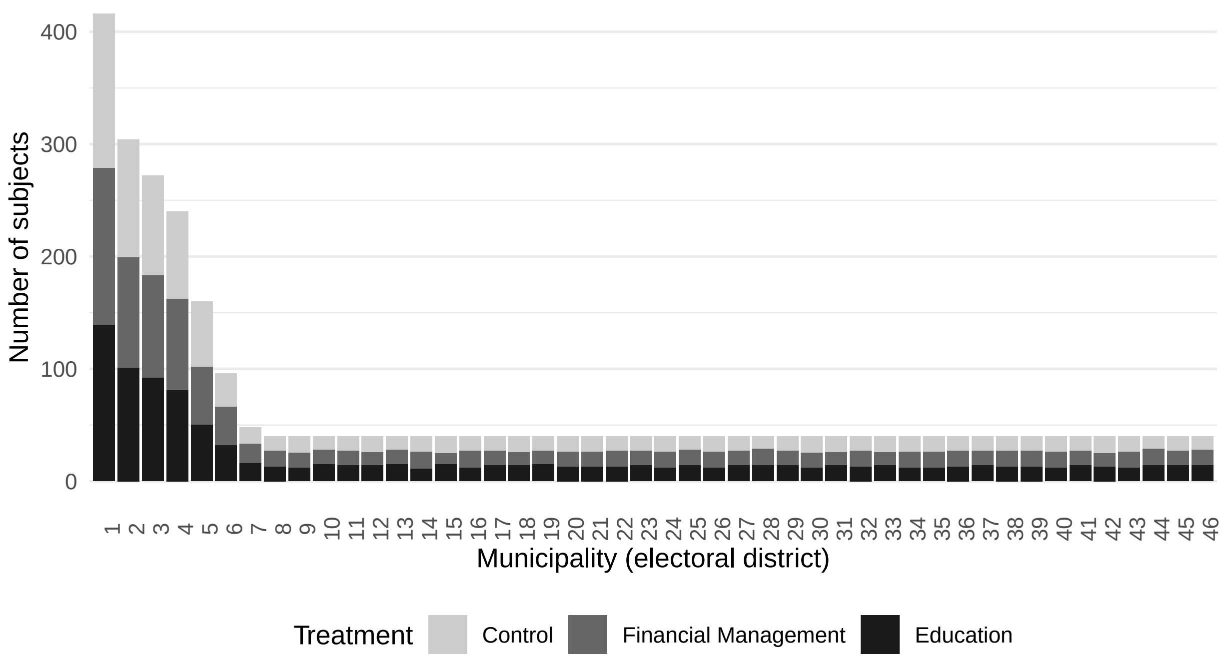


Figure A4: Sample allocation across districts/constituencies in Boas et al. (2019).

3. **Calculating** $MAEI_d$: This experiment presents a very straightforward calculation of the $MAEI_d$. Begin by assuming SUTVA (Assumption 1). Because the intervention would not happen in the absence of the experiment, we know that $|S_{01}^{cd}| = 0$. (In other words, no voters are being denied the intervention because of the experiment.) In an individually randomized experiment, $n_c = 1 \forall c$. This implies that $a_i(0) \in \{0, 1\} \forall c$ since each voter either votes for the marginal winner ($a_i(0) = 1$) or does not ($a_i(0) = 0$). Further, we know that $\sum_{c \in d} |S_{10}^{cd}|$ is simply the number of treated voters in each district, which is depicted by the two darker bar segments in Figure A4. Further, we know that n_d is the number of registered voters in each municipality. In the Brazilian case, this data is easily accessible from the Brazilian Tribunal Superior Eleitoral (TSE).

   The results of the $MAEI_d$ calculation for each district under SUTVA (potentially non-conservative) and assuming $E[a_i(0) = 0]$ for all treated voters (equivalent to $E[a_i(0) = 1]$ and conservative) is graphed in Figure A5. Under these assumptions, the $MAEI_d$ collapses to the number of treated voters over the number of registered voters.

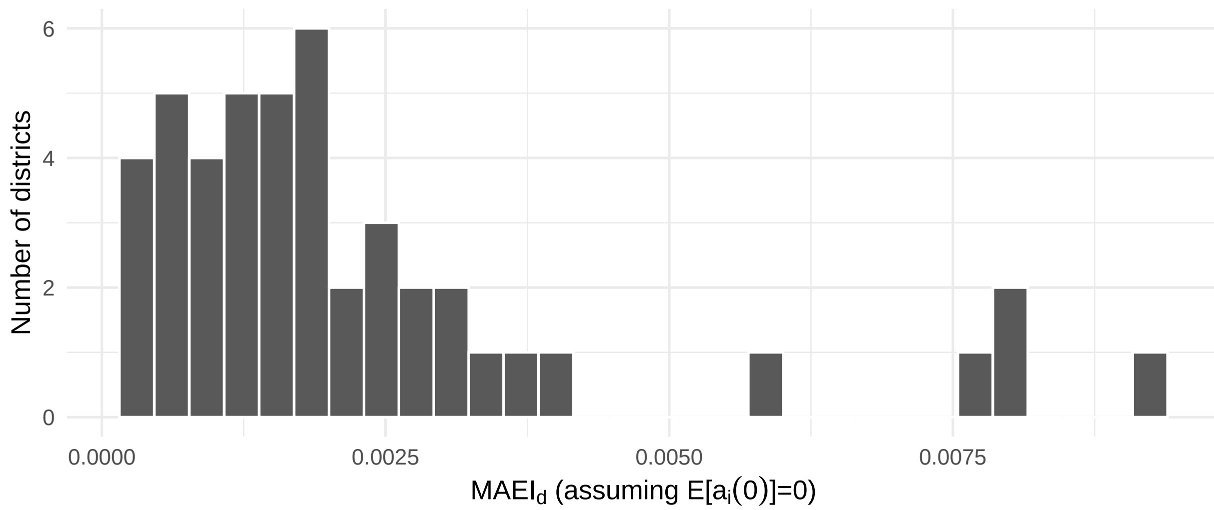


Figure A5: $MAEI_d$ for the 46 districts in Boas et al. (2019) under SUTVA and an assumptions that $E[a_i(0) = 0]$ for all treated voters.

Figure A6 depicts how the $MAEI_d$ changes under less conservative assumptions about $E[a_i(0)]$ (left) and more conservative assumptions about interference in the form of modest (relative to treatment density) between-cluster spillovers (right). Given that the PAP does not anticipate spillovers or interference, I use the otherwise conservative estimates of $MAEI_d$ from Figure A5 for the following analysis.

4. **Calculating** $\underline{\psi}_d$: I am unaware of systematic pre-election polling across municipalities leading up the Brazilian elections of 2016. To demonstrate the approach, I rely simply on predicting the mayoral margin of victory of the 2016 election from the margin of victory in the preceding 2012 election following the method in Appendix A7.3 with a single predictor: 2012 margin of victory. I fit the data on changes in margin of victory from 2004-2008 and 2008-2012 in the state of Pernambuco (both inside and outside the experimental municipalities). Figure A7 depicts the resultant estimates of $\underline{\psi}_d$ on the y-axis against the $2MAEI_d$ on the x-axis. Points (municipalities) below the 45° line do not pass the decision rule. Points (municipalities) above the 45° line do pass the rule. Using these estimates, 11/46 municipalities pass the decision rule.

   I encourage caution in interpreting Figure A7. A better predictive model would, in theory, be less conservative. The current estimates of $\underline{\psi_d}$ are penalized by the weak predictive model ($R^2 = 0.044$). The model is weakly predictive because of limited autocorrelation in margins of victory ($\sigma = 0.21$).

5. **Ex-post assessment** As depicted in Figure A8, I find that one of the mayoral races fell within $MAEI_d$. We certainly cannot ascertain from this finding that the treatment changed the election outcome. This election was decided by five votes. Nevertheless, it is useful to think about how the $MAEI_d$ compares to the ex-post distribution of election results. One further note is that if we were to consider municipal council races, the margin between the two marginal candidates is certainly closer than the distribution of the ψ_d in mayoral races.

A5.2 Application #2: Gerber and Green (2000)

Gerber and Green (2000) is a seminal study of voter mobilization that was conducted in New Haven, CT in the 1998 midterm elections.

1. **Electoral context:** Gerber and Green (2000) report the results of a voter mobilization campaign in New Haven, CT conducted during the November 3, 1998 midterm election. In this election, New Haven voters voted for

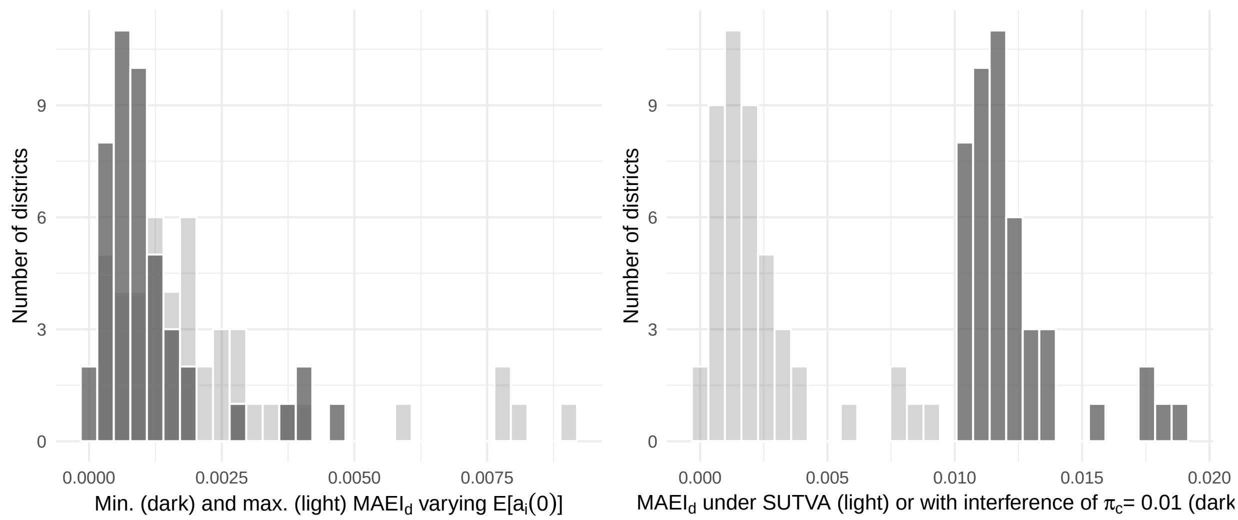


Figure A6: $MAEI_d$ for the 46 districts in Boas et al. (2019) varying $E[a_i(0)]$ (left) and allowing for between-cluster spillovers of magnitude $\pi_c = 1$ (right).

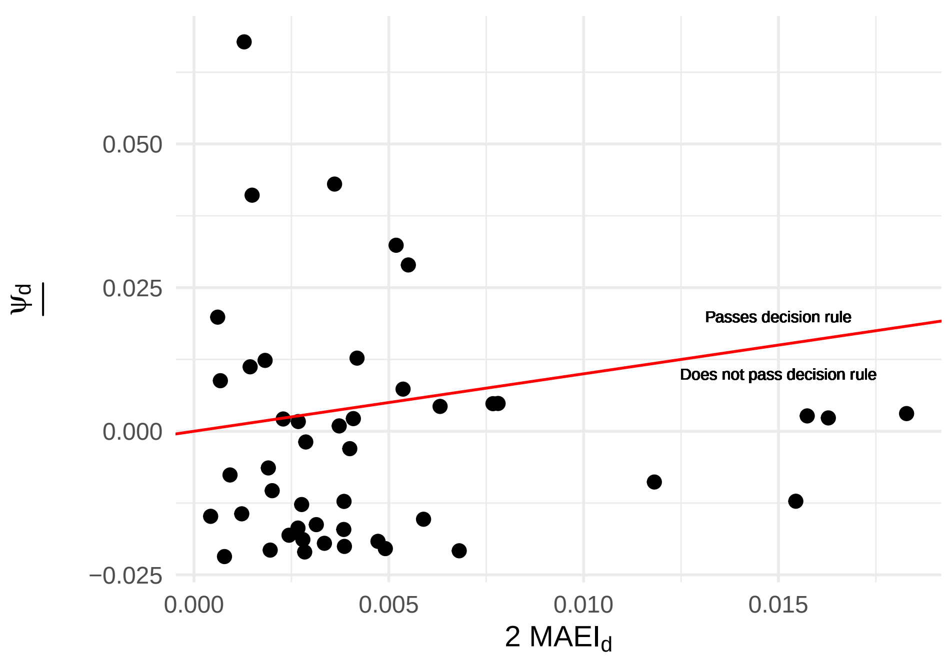


Figure A7: Visualization of the decision rule across the 46 experimental municipalities. Points below the 45° line do not pass the decision rule. Points above the 45° line do pass the rule.

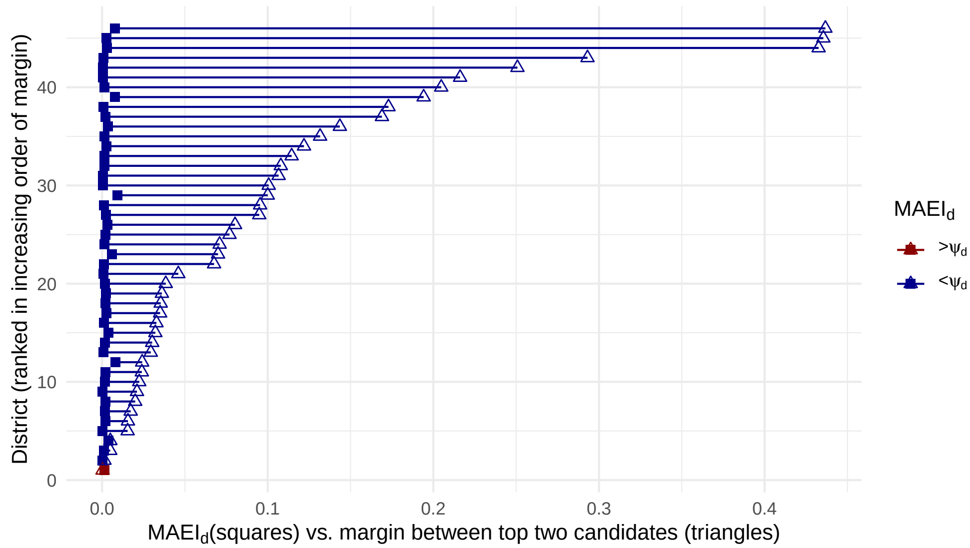


Figure A8: Comparison of $MAEI_d$ projections to the outcome of mayoral races in 2016. Here ψ_d corresponds to the actual (not the predicted margin of victory.)

several statewide offices (Governor, Secretary of State, Treasurer, Comptroller, Attorney General, and US Senate); US House (in district CT-3); state legislative seats (two state Senate races and seven state House races); one county race (New Haven County Sheriff); and one city race (for probate court judge). As a city, New Haven voters predominantly supported Democratic in 1998. Several races were uncontested including 1/2 state Senate races; 3/7 state House races; and the probate court race.

2. **Experimental overview** Gerber and Green (2000) randomly assigned voters in New Haven, Connecticut to various get out the vote treatments prior to the November 1998 (midterm) election. According to the Connecticut Secretary of State, there were 52,761 registered voters in New Haven as of November 3, 1998.[6] The researchers reduced the experimental sample by excluding individuals with PO boxes, households with more than two registered voters, and a university ward, for a total of 30,544 voters residing in 23,008 households.[7]

   Households were cluster-randomized to sixteen treatment arms that varied: (1) the number of direct mailings sent (0-3 mailings); (2) in-person canvassing (yes/no); and telephone canvassing (yes/no). Treatments were administered by researchers and graduate student assistants. In the absence of this research, voters would not have been contacted by mail, phone, or in-person canvass *by the research team for the purpose of research*. This does not preclude other interventions. I assume that the control group represents the "status quo" at least respect to the absence of interventions by the researchers. As such, I consider a control group of 10,800 voters and a treatment group of 18,580 voters.

3. **Calculating the $MAEI_d$:** The concurrent elections with different districts complicates the calculation of the $MAEI_d$. Consider two types of districts:

   - Districts that contain all experimental households. This includes all statewide races, the US House race (to the best of my knowledge), and the city and county races. For the purposes of this analysis, I consider the

[6] Data retrieved from `https://portal.ct.gov/SOTS/Election-Services/Statistics-and-Data/Statistics-and-Data`.

[7] These numbers come from a revised household-level replication dataset published by the authors in 2005. They are slightly higher than those reported in Gerber and Green (2000). No substantive conclusions below are affected by differences in these numbers.

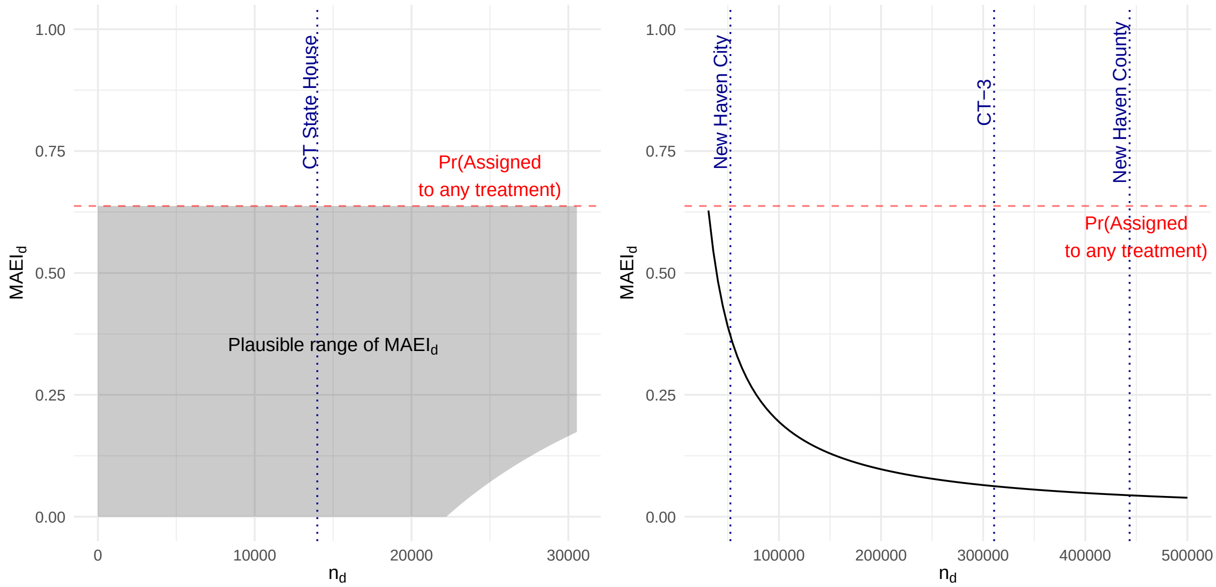


Figure A9: The left panel gives the bounds on the $MAEI_d$ for districts that do not contain all experimental households. The right panel gives the $MAEI_d$ as a function of n_d. Note that both plots assume SUTVA and that $E[a_i(0)] = 0$.

largest and smallest electorates of these districts: the statewide ($n_d = 1,806,750$) and city-wide electorate ($n_d = 52,761$).

- Districts that contain a subset of experimental households. Replication does not include the mapping to the relevant State Senate and State House districts (likely for privacy reasons). Nor is public registration compiled at this level. I roughly approximate the number of voters in State House and Senate districts as the number of registered voters statewide divided by the number of districts statewide. I consider the smaller districts – those for State House – with approximately $n_d = \frac{1{,}806{,}750}{133} = 13,585$ voters.

Given that clusters (households) were fully saturated (all voters were considered to be assigned to treatment), Assumption 2 is subsumed by SUTVA (Assumption 1). Given that this research was implemented by the researchers, $|S_{01}^{cd}| = 0$. In other words, no voters are being denied the intervention because of the experiment. We can construct conservative bounds (under SUTVA) by assuming $E[a_i(0)] = 0 \forall c$. In this case, the $MAEI_d$ is simply the number of treated voters as a fraction of the registered voters in the district $\frac{\sum_{c \in d} |S_{10}^{cd}|}{n_d}$. This calculation is straightforward when the experiment is entirely contained within a district (first case above).

When the experiment is not fully contained within a district, we cannot ascertain the $MAEI_d$ without further information relating experimental subjects to districts. We can, however, bound the plausible range of $MAEI_d$'s under the same assumptions. First, to bound the maximum, imagine that all subjects of a district were experimental subjects. Here, given random assignment of treatment, the proportion of voters assigned to treatment should be equivalent (in expectation) to the probability of assignment to (any) treatment. At a minimum, there may be no experimental subjects in a district. New Haven had 52,761 registered voters. It is plausible that any district smaller than the difference between the number of registered voters and the number of experimental subjects or 52,761-30,544 = 22,217 registered voters could be completely excluded from the experiment. Above that threshold, we can derive a minimum bound on the plausible $MAEI_d$ based on the probability of assignment to treatment and the number of registered voters in a district. Figure A9 proves a visualization. Note that these bounds are exceedingly wide in the absence of information about the relationship between voters and districts.

4. **Predicting** ψ_d: I am unable to predict many of the down-ballot races. However polls compiled by the Cook

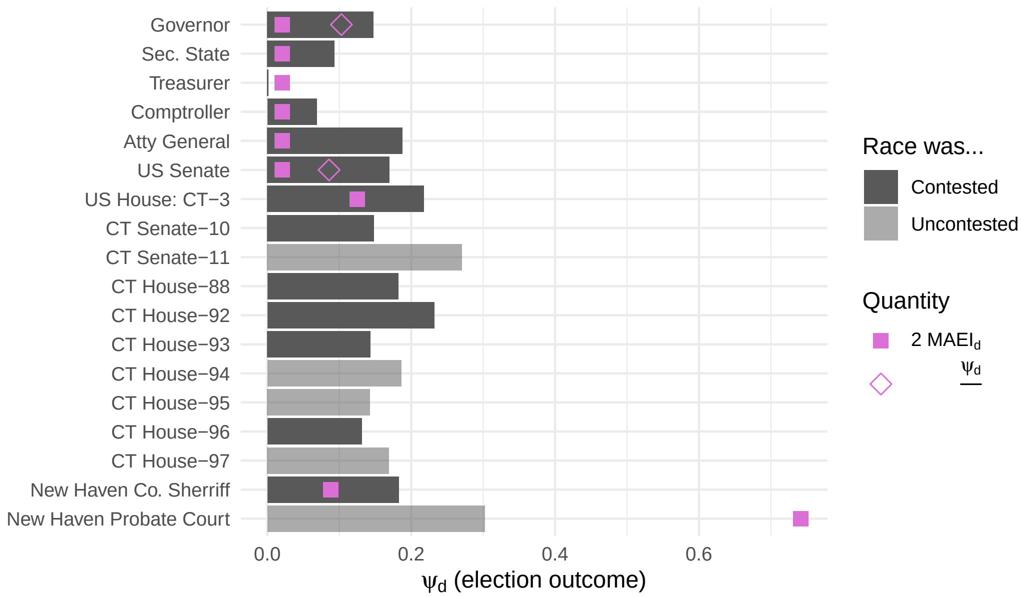


Figure A10: Comparison of estimable $MAEI_d$'s, $\underline{\psi}_d$'s, and electoral outcomes (ψ_d). See #3 and #4 for the issues that preclude estimation of some $MAEI_d$'s and $\underline{\psi}_d$'s.

Political Report allow for prediction of ψ_d for several state-level races. I report predictions from the last poll prior to the election in Table A6.

Race	Favored party	Survey	n	$\underline{\psi}_d$
Governor	R	Mason-Dixon PMR (Oct 9-12, 1998)	629	0.24
Senator	D	Mason-Dixon PMR (Oct 9-12, 1998)	629	0.20

Table A6: Estimates of $\underline{\psi}_d$ for two races for which polling data remains available.

Two notes are necessary. First, these estimates represent margin as a share of voters, not as a share of registered voters. Recall that ψ_d is constructed as a ratio to the number of registered voters. I normalize the rates in Table A6 by the 1997 elections (42.9%). Because 1997 was an off-cycle election, this is likely a conservative estimate (lower turnout than predicted in the midterm election in 1998). The points representing these $\underline{\psi}_d$ estimates in Figure A10 are equivalent to multiplying the predicted margin by this turnout rate. Second, several of the races are uncontested. In these cases, imputation of $\underline{\psi}_d = 1$ seems appropriate, particularly when there exists no turnout requirement. In Figure A10, I distinguish contested from uncontested races.

5. **Ex-Post Assessment**: Figure A10 compares (1) the election results (stated in terms of ψ_d, the margin of victory), by race; (2) the point estimable $MAEI_d$'s; and (3) the estimable $\underline{\psi}_d$'s. Two observations of note. First, both races for which we have all three quantities pass the decision rule. $2MAEI_d < \underline{\psi}_d$. Further, in the city race, the race was uncontested. Here $2MAEI_d$ presents less of a concern given the uncontested nature of the race. In contrast, the attorney general's race was very closes (decided by a mere 2,684 votes, or $\psi_d = 0.003$). This means that the intervention *could* have changed the outcome, though we can never know way.[8]

A5.3 Application #3: Bond et al. (2012)

Bond et al. (2012) is a very large-scale field experiment conducted in collaboration with Facebook during the 2010

[8]Given the modest estimates on turnout, it seems highly unlikely that this was the case in the absence of large, unmeasured persuasion effects.

midterm elections.

1. **Electoral Context**: In 2010, US midterm elections included election of 35 Senators, the entire House of Representatives, 34 Governors, and countless state and local officials. At the national and state levels, Republicans picked up many seats.

2. **Experimental Overview**: This experiment was presumably conducted at national scale. There are some vagaries due to the proprietary Facebook data. It included three treatment arms:

   - A *social message group* in which users saw a banner with a clickable "I voted" button at the top of their "Newsfeeds" ($n = 60,055,176$). This is consistent with Facebook's banner that a large number of users saw in 2008.
   - An *informational message group* without the button on their "Newsfeed" ($n = 611,044$).
   - A *control group* in which users did not see anything on the top of their "Newsfeed" ($n = 613,096$).

   An author interview with a researcher on this study indicates that in the absence of the experiment, all subjects would have seen the *social message group*. This implies that $\sum_{c \in d} |S_{10}^{cd} \cup S_{01}^{cd}| = 1,224,140$ (the sum of the subjects in the control and informational treatment arms). Obviously, this is a very small proportion ($\approx 2\%$) of the experimental sample.

   The US Census Bureau estimates that there were 137,263,000 registered voters in 2010.[9] As such, the experiment included 43.8% of registered voters with 0.9% assigned to one of the treatment conditions that existed *because of the experiment*.

3. **Calculating** $MAEI_d$: There are two important sources of ambiguity relevant to calculation of the $MAEI_d$. First, we do not know with certainty whether Facebook conducted the experiment across all possible US voters. The rate of treatment saturation is consistent with (outside Facebook) estimates of adult Facebook usage in 2010, so it is plausible that all states were included. Below, I make the simplifying assumption that Facebook intervened in all states/districts. Second, I do not know the distribution of subjects – or Facebook users – across the United States. To my knowledge, this data is not available. For the purposes of exposition, suppose that the proportion of registered voters in an arbitrary district, d, is $FB_d \in (0, 1]$.

   One aspect of treatment assignment is straightforward. The experiment did not use blocked random assignment and maintained constant probabilities of assignment to treatment across the experimental sample.[10] As above, the probability of assignment to treatment implies that $E[\frac{\sum_{c \in d} |S_{01}^{cd} \cup S_{10}^{cd}|}{n_d / FB_d}] = \frac{1,224,140 \times FB_d}{60,055,176} = 0.02 \times FB_d$.

   I consider two estimates of the $MAEI_d$, under SUTVA and with between-voter ("between cluster") spillovers. First, consider the calculation with $SUTVA$. In principle, it would be useful to know $E[a_i(0)]$ for the Facebook user population in the experimental sample. This may be approximated by data on "I voted" use in the 2006 election.

A5.4 Application #4: López-Moctezuma et al. (2021)

1. **Electoral Context**: The 1987 Filipino Constitution mandates that 20% of the House of Representatives (lower chamber) be composed by members of marginalized groups. Since 1995, this representation mandate has been fulfilled using closed-list proportional representation within a single national district to fill 58 seats of the 289-seat chamber. Seats are apportioned according to a Hare quota with a 2% threshold. Note that major parties are prohibited from running in these races that seek to provide representation to marginalized groups. There are many party lists that compete for the 58 seats. In the 2013 election studied by López-Moctezuma et al. (2021), 122 parties contested the party list contest. There were 52,006,910 registered voters in this election.[11]

[9] See `https://www.census.gov/data/tables/2010/demo/voting-and-registration/voting-registration-2010-election.html`.
[10] This point was confirmed by the author's interview with a researcher on the Bond et al. (2012) paper.
[11] National electoral outcome data comes from `https://en.wikipedia.org/wiki/2013_Philippine_general_election`.

Assignment	Treatments delivered "Business as usual" (Control)	 Deliberative town halls (Treatment)	$\|S_{10}\|$ (as assigned)
Control	24 barangays 178,635 RV	2 barangays 15,141 RV	
Treatment	2 barangays 15,542 RV	11 barangays 83,406 RV	98,948 RV
$\|S_{10}\|$ (as delivered)		98,547 RV	

Table A7: Non-compliance with treatment assignment. RV stands for registered voters.

Note that both of the parties with whom researchers collaborated were relatively small. However, improving the prospects of a small party can remove votes from a marginal seat winner, changing election outcomes. As such, I employ symmetric considerations regardless of party size.

2. **Experimental overview**: López-Moctezuma et al. (2021) randomize the campaign strategy of two party lists. The collaborations with each party list were conducted in different barangays (neighborhoods/villages). Treatment was assigned at the level of the barangay (the cluster) and the treatment is the use of deliberative town-hall meetings whereas control is "business as usual" campaigning. Per an interview with a researcher on this project, in the absence of the intervention, parties would have campaigned in treatment barangays in the "business as usual" fashion. Here, we can assume that there are no "if non-experiment assigned" communities, i.e. $|S_{01} = 0|$.

   One point worth discussion is that there was two-sided non-compliance in the administration of this intervention. Two barangays assigned to treatment were untreated and two barangays assigned to control were ultimately treated. The two-sided non-compliance occurred because researchers sought to assign additional barangays to treatment after the two treatment barangays did not participate in the debates. As such, two (very similar) quantities are relevant. While planning, the number of registered voters in treatment barangays would be the relevant measure of $|S_{10}| = 98,547$. During the administration, that number shifted to $|S_{10}| = 98,948$ when the replacement barangays were assigned. Fortunately these numbers are very similar.

3. **Calculating** $MAEI_d$: Here, there is a single district in which $n_d = 52,006,910$. Both parties were both small: in the preceding election, the larger party won 3% of the vote (as a share of valid votes) and the other party won 0% of the vote. We can approximate $E[a_c(0)] = 0$, without a large loss of accuracy. We know from above that $|S_{10} \cup S_{01}| = S_{10}$ given the definition of treatment. As such, $MAEI_d = \frac{|S_{10}|}{n_d}$. With either measure of $MAEI_d$, this simplifies to approximately 0.0019.

4. **Predicting** $\underline{\psi_d}$: In a proportional representation race with 58 seats and >100 party lists, it is very difficult to predict ψ_d. Analytically, we know that the upper bound on ψ_d is $\frac{1}{58} = 0.017$, which is substantially larger than the $2MAEI_d$ calculated above. However, this is a strict upper bound. I lack the tools to predict this quantity, though it is evident that even if $\psi_d \sim U[0, \frac{1}{58}]$ an admittedly very rough approximation, $F_{\psi_d}(2MAEI_d) = 0.22$, outside the decision rule. There may be much better ways to calculate ψ_d with more electoral information.

5. **Ex-Post Assessment**: The realized $\psi_d = 0.0002$ was calculated from the marginal party earning 1 seat (Append) and the marginal party earning 0 seats (Alif). This margin is smaller than between the marginal party winning 3 seats and the marginal party winning 2 seats as well as the marginal party winning 2 seats and the marginal party winning 1 seat. This is clearly smaller than $2MAEI_d$. Note that of course that this does not indicate that the experiment changed the aggregate outcome, only that it had the potential to.

A6 Existing Information Experiments

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

Table A8 describes studies in the framework described in this paper. Thirteen of the 14 studies intervene in multiple races (districts). I focus on calculating the *average* $MAEI_d$ across districts. The *average* $MAEI_d$ is an abstraction from the decision rule described in this paper. However, for the purposes of examining the literature, it does serve as a measure of the variation across studies on this metric. I am only able to estimate the $MAEI_d$ in six of 14 studies, varying $E[a_c(0)]$ from its minimum of 0.5 (for all c) to its maximum of 1 (for all c). I present these estimates in Figure 1. The graph suggests that the degree to which existing information experiments could have moved electoral outcomes varies widely. Recall that these estimates in isolation cannot assess whether an intervention was consistent with the decision rule advocated here because I lack data on the predicted margin of victory. Nevertheless, any $MAEI_d > 0.5$ will not pass the decision rule, regardless of the predicted *ex-ante* margin of victory. One immediate observation from Figure 1 is that cluster-assigned information treatments appear to be assigned at a very high density within districts.

Article	Country	Mapping to Framework	Calculation Details	$MAEI_d$ Est.	$\|D\|$
Adida et al. (2017)	Benin	d : Commune* c: Village (or urban quarters) j: Individual	Treatment (five variants) was assigned to 195 of 1498 villages. The saturation (density) of treatment in a village varies by treatment arm (below 100% in all villages) and village population is unclear without data on cluster size. Because of the noted distinction in the *de-jure* vs. *de-facto* characterization of parliamentary electoral districts, more information needed to clarify n_d.	–	30
Arias et al. (2019)	Mexico	d : Municipality c : Precinct j : Individual	At most $\frac{1}{3}$ of precincts per municipality were sampled, albeit non-randomly. Treatment was assigned to 200 households in each of 400 precincts (T1-T4). Precincts contain a maximum of 1,750 registered voters. Non-random sampling of precincts prevents calculation of $MAEI_d$.	–	26
Banerjee et al. (2011)	India	d : State leg. district c : Polling station j : Individual	20 treated polling stations and average of 57.5 control polling stations per district. All households were treated.	$[0.129, 0.258]$	10
Bhandari et al. (2020)	Senegal	d: Department c: Village d: Individual	9 individuals were sampled per village. 450 villages were (non-randomly) sampled from the 859 villages in the 5 experimental departments. 375 villages received some treatment (non pure-control). Without further information on the distribution of villages (experimental and non-experimental) and population by district, the $MAEI_d$ cannot be calculated.	–	5
Boas et al. (2019)	Brazil	d : Municipality 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
Chong et al. (2015)	Mexico	d: Municipality c : Precinct j : $Individual$	450 of 2360 precincts were treated (selected randomly). No information is provided on saturation within precinct so I assume all voters were treated.	$[0.095, 0.191]$	12

Article	Country	Mapping to Framework	Calculation Details	$MAEI_d$ Est.	$\|D\|$
Cruz et al. (2019)	Philippines	d : Municipality c : Village j: Individual	All households in 104 treatment villages (T1 or T2) across 7 municipalities were visited. Each municipality has "20-25 villages." I assume 25 villages/municipality and that the experimental villages were randomly sampled.	$[0.279, 0.594]$	7
Cruz et al. (2018)	Philippines	d : Municipality c : Village j: Individual	All households in 142 treatment villages in 12 municipalities were visited. The average number of villages/municipality not reported. I assume 25 villages/municipality per Cruz et al. (2019) (which is consistent with 284 villages in the experimental sample). Villages were randomly sampled from the municipality.	$[0.237, 0.473]$	12
de Figueiredo et al. (2011)	Brazil	d : Municipality c : Precinct j : Individual	$\approx$ All households were visited with flyers in 200 treatment (T1 or T2) precincts of 1,759 precincts in the municipality. The precincts were selected randomly subject to a set of constraints.	$[0.057, 0.114]$	1
George et al. (2018)	India	d : Assembly constituency c : Village j : Individual	The intervention treated 500,000 voters (T1-T4) in 1,591 villages. Villages have $\approx 1,200$ registered voters, so saturation rate in treatment villages was averaged 26%. Non-random sampling of villages within constituencies prevents estimation of $MAEI_d$.	–	38
Humphreys and Weinstein (2012)	Uganda	d : Parliamentary constituency c : Polling station j: Individual	2 polling stations in selected constituencies and all households visited with flyers. Number of polling stations/constituency not reported so $MAEI_d$ cannot be calculated. The total number of constituencies where experiment occurred (known to be <147) is not reported.	–	–
Lierl and Holmlund (2019)	Burkina Faso	d: Village* c: Individual j: Individual	12 individuals were assigned to treatment (T or placebo) per village. Information about village population (n_d) is not reported.	$\frac{12}{n_d}$	146
Sircar and Chauchard (2019)	India	d : Assembly Constituency c: Polling booth area j : Individual	16 polling booth areas per precinct assigned to treatment (T1 or T2) with $\frac{2}{3}$ of households in each polling booth area assigned to receive flyer. While selection of experimental polling booths is random, the total number of polling booths per constituency is not reported so $MAEI_d$ cannot be calculated	–	25

Table A8: 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.

A7 Supporting Information for Empirical Illustration

A7.1 Data and Data Sources

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

- Colorado:
  - Precinct-level electoral returns and voter registration from Colorado Secretary of State `https://www.coloradosos.gov/pubs/elections/Results/archive2000.html` and `https://www.sos.state.co.us/pubs/elections/VoterRegNumbers/VoterRegNumbers.html`.
  - 2018 House of Representative seat predictions from The Crosstab (downloaded in 2019).

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

Table A9: Summary statistics on State and US House districts in terms of registered voters and precincts. Note that past electoral data from 2012, 2014, and 2016 is also collected for use in prediction.

A7.2 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 A10.

A7.3 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 estimate the predictive distribution of each ψ_d in (effectively) two-party races, following the steps:

1. Estimate a model of the form: $y_d = f(\boldsymbol{\beta}\mathbf{X}_d)$, where y_d is the margin of victory for the Democratic party (as a share of registered voters), $\mathbf{X}_d$ is a matrix of predictors. Note that the unit of analysis is the district, d.
2. Calculate residuals from the model, $\widehat{\epsilon}_d = y_d - \widehat{y_d}$ for each district d. Denote the distribution of residuals as $f_{\widehat{\epsilon}}$.
3. Generate many draws from the joint distribution of $\widehat{\boldsymbol{\beta}}$. For each draw:
   (a) Estimate $\widehat{\psi_d}$ from the model.
   (b) Randomly sample $x_d \sim f_{\widehat{\epsilon}}$ (independently across d) and calculate $\widehat{\psi_d} + x_d$.
4. These estimates form the empirical distribution $f(\psi_d|\widehat{\boldsymbol{\theta}})$ for each district, d.

A7.4 Maximum Number of Treated Clusters in Simulation

Figure A11 conducts a similar analysis to that conducted in Figure 3 except treating the precinct as the unit of randomization. Here, we assume that entire precincts are assigned to treatment. Given the election predictions for both Colorado State House and US House races, this simulation evaluates the expected maximum number and proportion of precincts could be assigned to treatment under the present decision rule. This assumes no between-cluster interference.

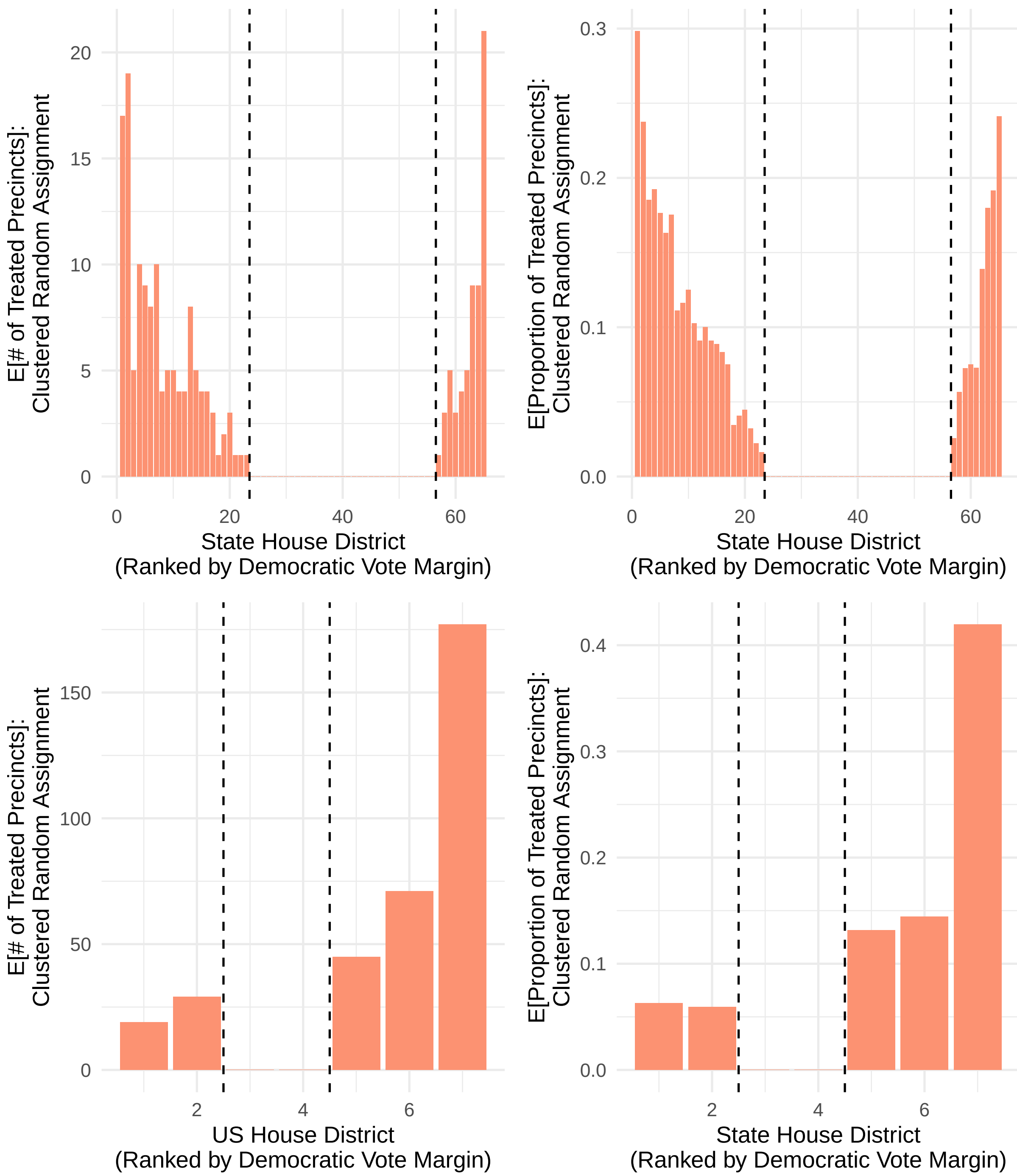


Figure A11: Maximum number of precincts (left) or and precincts as a proportion of all precincts (right) that can be assigned to treatment under decision rule. The expectation is evaluated over simulated assignments of treatment to precincts with heterogeneous numbers of registered voters.

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

Table A10: 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 $|S_{01}| = 0$.

Supplementary Appendix: References

Adida, C., J. Gottlieb, E. Kramon, and G. McClendon (2017). Reducing or reinforcing in-group preferences? an experiment on information and ethic voting. *Quarterly Journal of Political Science 12*(4), 437–477.

American Political Science Association (2020, April). Principles and guidance for human subjects research. Available at https://tinyurl.com/y5vm6cem.

Arias, E., H. Larreguy, J. Marshall, and P. Querubin (2019, February). 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`.

Asunka, J., S. Brierly, M. Golden, E. Kramon, and G. Ofosu (2017). Electoral fraud or violence: The effect of observers on party manipulation strategies. *British Journal of Political Science 49*(1), 129–151.

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.

Banerjee, A., S. Kumar, R. Pande, and F. Su (2011, November). 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, E., R. Davis, and A. Kern (2020). The democratic limits of political experiments. *Politics, Philosophy, and Economics 19*(4), 321–342.

Bhandari, A., H. Larreguy, and J. Marshall (2020). Able and mostly willing: An empirical anatomy of information's effect on voter-driven accountability in senegal. *American Journal of Political Science Forthcoming*. Available at `https://scholar.harvard.edu/files/jmarshall/files/accountability_senegal_paper_v5.pdf`.

Bidwell, K., K. Casey, and R. Glennerster (2020). Debates: Voting and expenditure responses to political communication. *Journal of Political Economy 128*(8), 2880–2924.

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.

Boatright, R. G., D. P. Green, and M. J. Malbin (2006). Does publicizing a tax credit for political contributions increase its use?: Results from a randomized field experiment. *American Politics Research 34*(5), 563–582.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler (2012). A 61-million-person experiment in social influence and political mobilization. *Nature 489*, 295–298.

Buntaine, M. T., R. Jablonski, D. L. Nielson, and P. 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, E. (2020, 89-93). Field experiments and behavioral theories: Science and ethics. *PS Political Science and Politics* (53), 1.

Casey, K., A. Kamara, and N. Meriggi (2021). An experiment in candidate selection. *American Economic Review Forthcoming*.

Chong, A., A. de la O, D. Karlan, and L. 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, C., P. Keefer, and J. Labonne (2018, September). 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, C., P. Keefer, J. Labonne, and F. Trebbi (2019, June). 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.

Davis, J. and K. Michelitch (2021). Field experiments: Thinking through identity and positionality. *PS Political Science and Politics Forthcoming*.

de Figueiredo, M. F., F. D. Hidalgo, and Y. Kasahara (2011, February). 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, S. (2018). Subjects and scholars' views on the ethics of political science field experiments. *Perspectives on Politics 16*(3), 739–750.

Dunning, T., G. Grossman, M. Humphreys, S. D. Hyde, C. McIntosh, and G. Nellis (Eds.) (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.

George, S., S. Gupta, and Y. Neggers (2018, November). 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, A. S. and D. P. Green (1999). Does canvassing increase voter turnout? a field experiment. *Proceedings of the National Academy of Sciences 96*(14), 10939–10942.

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.

Gubler, J. R. and J. S. Selway (2016). Considering the political consequences of comparative politics experiments. In S. Desposato (Ed.), *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*. New York: Routledge.

Humphreys, M. (2015). Reflections on the ethics of social experimentation. *Journal of Globalization and Development 6*(1), 87–112.

Humphreys, M. and J. M. Weinstein (2012, March). Policing politicians: Citizen empowerment and political accountability in uganda - preliminary analysis. IGC Working Paper S-5021-UGA-1.

Jha, S. and M. Shayo (2019). Valuing peace: The effects of financial market exposure on votes and political attitudes. *Econometrica 87*(5), 1561–1588.

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.

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

López-Moctezuma, G., L. Wantchekon, D. Rubenson, T. Fujiwara, and C. P. Lero (2021). Policy deliberation and voter persuasion: Experimental evidence from an election in the philippines. *American Journal of Political Science Forthcoming*.

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.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The belmont report: Ethical principles and guidelines for the protection of human subjects of research. Technical report, U.S. Department of Health and Human Services.

Shineman, V. (2020, September). Restoring rights, restoring trust: Evidence that reversing felony disenfranchisement penalties increases both trust and cooperation with governmetn. Working paper, available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3272694`.

Sircar, N. and S. Chauchard (2019). *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*, Chapter Dilemmas and Challenges of Citizen Information Campaigns: Lessons from a Failed Experiment in India, pp. 287–311. Number 10. New York: Cambridge University Press.

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

Teele, D. L. (2019, June). Virtual consent: The bronze standard for experimental ethics. In preparation for *Advances in Experimental Methodology* volume.

Wantchekon, L. (2003). Clientelism and voting behavior: Evidence from a field experiment in benin. *World Politics 55*(3), 399–422.

Whitfield, G. (2019). Toward a separate ethics of political field experiments. *Political Research Quarterly* (1-12).

Zimmerman, B. (2016). Information and power: Ethical considerations of political information experiments. In S. Desposato (Ed.), *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*, pp. 183–197. New York: Taylor and Francis.