Corruption information and vote share: A meta-analysis and lessons for survey experiments*

Trevor Incerti†

November 1, 2019

Abstract

Do voters hold politicians accountable for corruption? Field experiments that provide voters with information about corrupt acts of politicians then monitor vote choice are now standard. Similarly, vote-choice survey experiments commonly inform respondents of corrupt acts of hypothetical candidates. Meta-analysis demonstrates that the aggregate treatment effect of corruption information on vote share in field experiments is approximately zero. By contrast, corrupt candidates are punished by approximately 32-34 percentage points across survey experiments. This suggests that while vote-choice survey experiments may identify the effect of informational treatments in hypothetical scenarios, the point estimates they provide may not be representative of real-world voting behavior. I explore publication bias, social desirability bias, and differences in experimental design and context as potential explanations for this discrepancy. Finally, I suggest examining the probability of voting for specific theoretically motivated candidates in conjoint experiments when researchers have priors about the conditions that shape voter decision-making.

*I am extremely grateful to Peter Aronow, Alexander Coppock, Angèle Delevoye, Devin Incerti, Joshua Kalla, Daniel Mattingly, Gautam Nair, Susan Rose-Ackerman, Frances Rosenbluth, Radha Sarkar, and Fredrik Sävje; participants of the 2019 APSA Corruption and Electoral Behavior Panel; participants at the Yale ISPS Experiments Workshop; the Yale Casual [sic] Inference Lab; and three anonymous reviewers for invaluable feedback and suggestions. Any and all errors are my own.

†PhD Student in the Department of Political Science, Yale University. trevor.incerti@yale.edu.

1 Introduction

Competitive elections create a system whereby voters can hold policy makers accountable for their actions. This mechanism should make politicians hesitant to engage in malfeasance such as blatant acts of corruption. Increases in public information regarding corruption should therefore decrease levels of corruption in government, as voters armed with information should expel corrupt politicians (Gray and Kaufman 1998; Kolstad and Wiig 2009; Rose-Ackerman and Palifka 2016). However, this theoretical prediction is undermined by the observation that well-informed voters continue to vote corrupt politicians into office in many democratic states. Political scientists and economists have therefore turned to experimental methods to test the causal effect of learning about politician corruption on vote choice.

Numerous experiments have examined whether providing voters with information about the corrupt acts of politicians decreases their re-election rates. These papers often suggest that there is little consensus on how voters respond to information about corrupt politicians (Arias, Larreguy, Marshall and Querubin 2018; Botero, Cornejo, Gamboa, Pavao and Nickerson 2015; Buntaine, Jablonski, Nielson and Pickering 2018; De Vries and Solaz 2017; Klašnja, Lupu and Tucker 2017; Solaz, De Vries and de Geus 2018). Others indicate that experiments have provided us with evidence that voters strongly punish individual politicians involved in malfeasance (Chong, De La O, Karlan and Wantchekon 2014; Weitz-Shapiro and Winters 2017; Winters and Weitz-Shapiro 2015,1).

By contrast, meta-analysis suggests that: (1) In aggregate, the effect of providing information about incumbent corruption on incumbent vote share in field experiments is approximately zero, and (2) corrupt candidates are punished by respondents by approximately 32-34 percentage points across survey experiments. I also examine mechanisms that may give rise to this discrepancy. I do not find systematic evidence of publication bias. I discuss the possibility that social desirability bias may lead survey experiments to capture anti-corruption norms rather than realistic voter behavior. Field and survey experiments also may be measuring different causal estimands due to differences in context and survey design. Conjoint experiments attempt to alleviate some of these issues, but are often analyzed in ways that may fail to illuminate the most substantively important comparisons. I suggest examining the probability of voting for candidates with specific combinations of attributes in conjoint experiments when researchers have priors about the conditions that shape voter decision-making, and using classification trees to illuminate these conditions when they do not.

2 Corruption information and electoral accountability

Experimental support for the hypothesis that providing voters with information about politicians' corrupt acts decreases their re-election rates is mixed. Field experiments have provided some causal evidence that informing voters of candidate corruption has negative (but generally small) effects on candidate vote-share. This information has been provided by: randomized financial audits (Ferraz and Finan 2008), fliers revealing corrupt actions of politicians (Chong et al. 2014; De Figueiredo, Hidalgo and Kasahara 2011), and SMS messages (Buntaine et al. 2018). However, near-zero and null findings are also prevalent, and the negative and significant effects reported above sometimes only manifest in particular subgroups. Banerjee, Green, Green and Pande (2010) primed voters in rural India not to vote for corrupt candidates, and Banerjee, Kumar, Pande and Su (2011) provided information on politicians' spending discrepancies, with both studies finding near-zero and null effects on vote share. Boas, Hidalgo and Melo (2018) similarly find zero and null effects from distributing fliers in Brazil. Finally, Arias et al. (2018); Arias, Larreguy, Marshall and Querubin (2019) find that providing Mexican voters with information (fliers) about mayoral corruption actually *increased* incumbent party vote share by 3%.[1]

By contrast, online survey experiments consistently show large negative effects from information treatments on vote share for hypothetical candidates. These experiments often

[1]The authors theorize that this average effect stems from levels of reported malfeasance actually being lower than voters' no-information expectations of corruption.

manipulate moderating factors other than information provision (e.g. quality of information, source of information, partisanship, whether corruption brings economic benefits to constituents, etc.), but even so systematically show negative treatment effects (Anduiza, Gallego and Muñoz 2013; Avenburg 2019; Banerjee, Green, McManus and Pande 2014; Boas, Hidalgo and Melo 2018; Breitenstein 2019; Eggers, Vivyan and Wagner 2018; Franchino and Zucchini 2015; Klašnja and Tucker 2013; Klašnja, Lupu and Tucker 2017; Mares and Visconti 2019; Vera 2019; Weitz-Shapiro and Winters 2017; Winters and Weitz-Shapiro 2013,1,1,1). These experiments have historically taken the form of single treatment arm or multiple arm factorial vignettes, but more recently have tended toward conjoint experiments (Agerberg 2019; Breitenstein 2019; Chauchard, Klašnja and Harish 2019; Franchino and Zucchini 2015; Klašnja, Lupu and Tucker 2017; Mares and Visconti 2019).

Boas, Hidalgo and Melo (2018) find differential results in a pair of field and survey experiments conducted in Brazil—zero and null in field, large and negative in survey. They argue that this may reflect that norms against malfeasance in Brazil do not translate into action in real life. Boas, Hidalgo and Melo (2018) argue that "differences in research design are unlikely to account for much of the difference in effect size" and point to features specific to Brazil in their explanation of this discrepancy, namely lower salience of corruption to Brazilian voters in municipal elections and the strong effects of dynastic politics. However, meta-analysis demonstrates that this is not only the case for Boas, Hidalgo and Melo (2018)'s experiments in Brazil, but extends across a systematic review of all countries and studies conducted to date. This suggests that the discrepancy between field and survey experimental findings is driven by differences in research design, rather than features specific to Brazilian voters.

Lab experiments that reveal corrupt actions of politicians to fellow players, then measure vote choice also show large negative treatment effects (see Figure A.1). While there are not enough lab experiments examining whether the provision of corruption information impacts vote choice to conduct a formal meta-analysis (Arvate and Mittlaender 2017; Azfar and

Nelson 2007; Rundquist, Strom and Peters 1977; Solaz, De Vries and de Geus 2018), this discrepancy is worth noting as previous examinations of lab-field correspondence have found evidence of general replicability (Camerer 2011; Coppock and Green 2015).

3 Research Design and Methods

3.1 Selection criteria

I followed standard practices to locate the experiments included in the meta-analysis. This included following citation chains and searches of data bases using a variety of relevant terms ("corruption experiment," "corruption field experiment," "corruption survey experiment," "corruption factorial", "corruption candidate choice", "corruption conjoint", "corruption, vote, experiment", and "corruption vignette"). Papers from any discipline are eligible for inclusion, but in practice stem only from economics and political science. Both published articles and working papers are included so as to ensure the meta-analysis is not biased towards published results. In total, I located 10 field experiments from 8 papers, and 18 survey experiments from 16 papers.

Field experiments are included if researchers randomly assigned information regarding incumbent corruption to voters, then measured corresponding voting outcomes. This therefore excludes experiments that randomly assign corruption information, but use favorability ratings or other metrics rather than actual vote share as their dependent variable. I include one natural experiment, Ferraz and Finan (2008), as random assignment was conducted by the Brazilian government. Effects reported in the meta-analysis come from information treatments on the entire sample of study only, not subgroup or interactive effects that reveal the largest treatment effects.

For survey experiments, studies must test a no-information control group versus a corruption information treatment group and measure vote choice for a hypothetical candidate. This necessarily excludes studies that compare one type of information provision (e.g. source) to another and the control group is one type of information rather than no information, or where the politician is always known to be corrupt (Anduiza, Gallego and Muñoz 2013; Botero et al. 2015; Konstantinidis and Xezonakis 2013; Muñoz, Anduiza and Gallego 2012; Rundquist, Strom and Peters 1977; Weschle 2016). In many cases, studies have multiple corruption treatments (e.g. high quality information vs. low quality information, co-partisan vs. opposition party, etc.). In these cases, I replicate the studies and code corruption as a binary treatment (0 = clean, 1 = corrupt) where *all* treatment arms that provide corruption information are combined into a single treatment. Studies that use non-binary vote choices are rescaled into a binary vote choice.[2]

3.2 Included studies

A list of all papers - disaggregated by field and survey experiments - that meet the criteria outlined above are provided in Table 1 and Table 2. A list of lab experiments (4 total) can also be found in and Table A.1, although these studies are not included in the meta-analysis. A list of excluded studies with justification for their exclusion can be found in Table A.2.

Table 1: Field experiments

Study	Country	Treatment
Arias et al. (2018)	Mexico	Fliers
Banerjee et al. (2010)	India	Newspaper
Banerjee et al. (2011)	India	Canvas/Newspaper
Boas, Hidalgo and Melo (2018)	Brazil	Fliers
Buntaine et al. (2018)	Ghana	SMS
Chong et al. (2014)	Mexico	Fliers
De Figueiredo, Hidalgo and Kasahara (2011)	Brazil	Fliers
Ferraz and Finan (2008)	Brazil	Audits

[2]For example, a 1-4 scale is recoded so that 1 or 2 is equal to no vote, and 3 or 4 is equal to a vote.

Table 2: Survey experiments

Study	Country	Type of survey
Agerberg (2019)	Spain	Conjoint
Avenburg (2019)	Brazil	Vignette
Banerjee et al. (2014)	India	Vignette
Breitenstein (2019)	Spain	Conjoint
Boas, Hidalgo and Melo (2018)	Brazil	Vignette
Chauchard, Klašnja and Harish (2019)	India	Conjoint
Eggers, Vivyan and Wagner (2018)	UK	Conjoint
Franchino and Zucchini (2015)	Italy	Conjoint
Klašnja and Tucker (2013)	Sweden	Vignette
Klašnja and Tucker (2013)	Moldova	Vignette
Klašnja, Lupu and Tucker (2017)	Argentina	Conjoint
Klašnja, Lupu and Tucker (2017)	Chile	Conjoint
Klašnja, Lupu and Tucker (2017)	Uruguay	Conjoint
Mares and Visconti (2019)	Romania	Conjoint
Vera (2019)	Peru	Vignette
Winters and Weitz-Shapiro (2013)	Brazil	Vignette
Winters and Weitz-Shapiro (2016)	Brazil	Vignette
Weitz-Shapiro and Winters (2017)	Brazil	Vignette
Winters and Weitz-Shapiro (2018)	Argentina	Vignette

3.3 Additional selection comments

Additional justification for the inclusion or exclusion of certain studies, as well as coding and/or replication choices may be warranted in some cases. Despite often being considered a form of corruption (Rose-Ackerman and Palifka 2016), I exclude electoral fraud experiments as whether vote buying constitutes clientelism or corruption is a matter of debate (Stokes, Dunning, Nazareno and Brusco 2013). The field experiment conducted by Banerjee et al. (2010) is included. However, the authors treated voters with a campaign not to vote for corrupt candidates in general, but did not provide voters with information on which candidates were corrupt. Similarly, the field experiment conducted by Banerjee et al. (2011) is included, but their treatment provided information on politicians' spending discrepancies, which may imply corruption but is not as direct as other types of information provision. Excluding these studies from the meta-analysis, the point estimates remain approximately zero percentage points using random effects estimation, and are approximately -1 percentage point using fixed effects estimation (see Figure A.2 and Table A.6).

With respect to survey experiments, Chauchard, Klašnja and Harish (2019) include two treatments, wealth accumulation and whether the wealth accumulation was illegal. The effect reported here is the illegal treatment only. This is likely a conservative estimate, as the true effect is a combination of illegality and wealth accumulation. Winters and Weitz-Shapiro (2016) and Weitz-Shapiro and Winters (2017) report results from the same survey experiment, as do Winters and Weitz-Shapiro (2013) and Winters and Weitz-Shapiro (2015). Each of these results are therefore only reported once. The survey experiment in De Figueiredo, Hidalgo and Kasahara (2011) is excluded from the analysis as it does not use hypothetical candidates, but instead asks voters if they would have changed their actual voting behavior in response to receiving corruption information. This study has a slightly positive and null finding. The overall results are not sensitive to the inclusion of this estimate (see Figure A.3).

4 Results

Survey experiments estimate much larger negative treatment effects of providing information about corruption to voters relative to field experiments. In fact, the field-experimental results in Figure 1 reveal a point estimate of approximately zero and suggest that we cannot reject the null hypothesis of no treatment effect. By contrast, Figure 2 shows that corrupt candidates are punished by respondents by approximately 34 percentage points in survey experiments based on fixed effects meta-analysis and 32 percentage points using random effects meta-analysis. Of the 18 survey experiments, only one shows a null effect (Klašnja and Tucker 2013), while all others are negative and significantly different from zero at conventional levels.

Examining all studies together, a test for heterogeneity by type of experiment (field or survey) reveals that up to 68% of the total heterogeneity across studies can be accounted for by a dummy variable for type of experiment (0 = field, 1 = survey). This dummy variable has a significant association with the effectiveness of the information treatment at the 1% significance level. In fact, with this dummy variable included, the overall estimate across studies is approximately 0, while the point estimate of the survey dummy is -0.32.[3] This implies that the predicted treatment effect across experiments is not significantly different from zero when an indicator for type of experiment is included in the model. In other words, the majority of the heterogeneity in findings is accounted for by the type of experiment conducted.

[3]Using a mixed effects model with a survey experiment moderator (see Table A.5). With Banerjee et al. (2010) and Banerjee et al. (2011) excluded from the model, the point estimate of the survey dummy is 0.31 and the heterogenity accounted for by the survey experiment moderator is 64% (see Table A.8 and Table A.7).

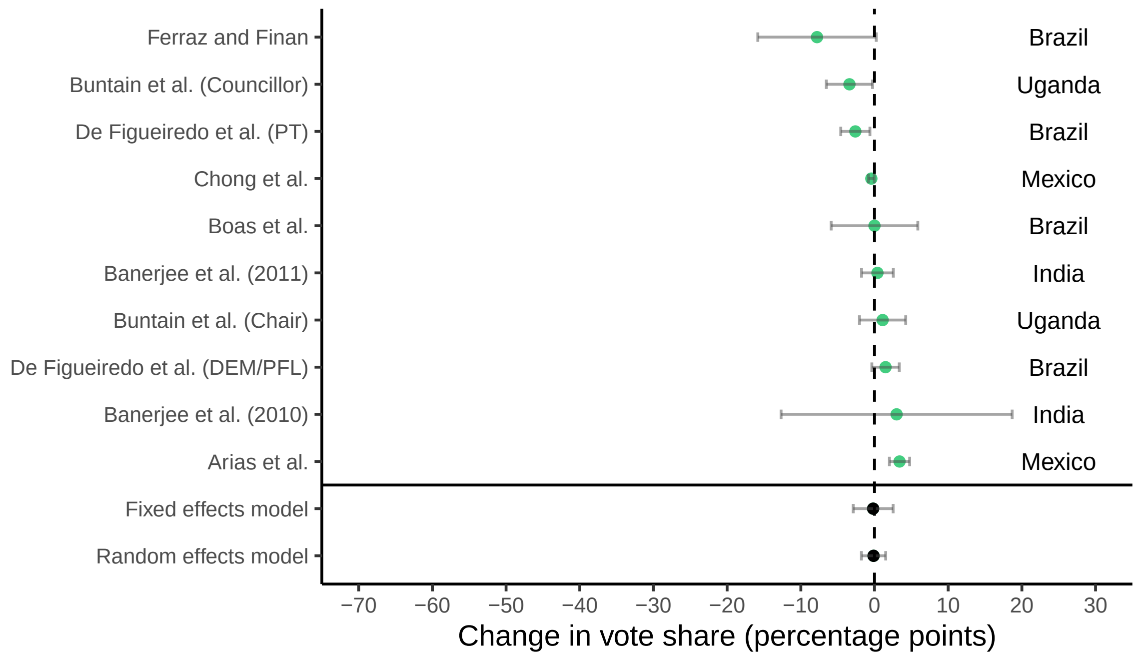

Figure 1: Field experiments: Average treatment effect of corruption information on incumbent vote share and 95% confidence intervals

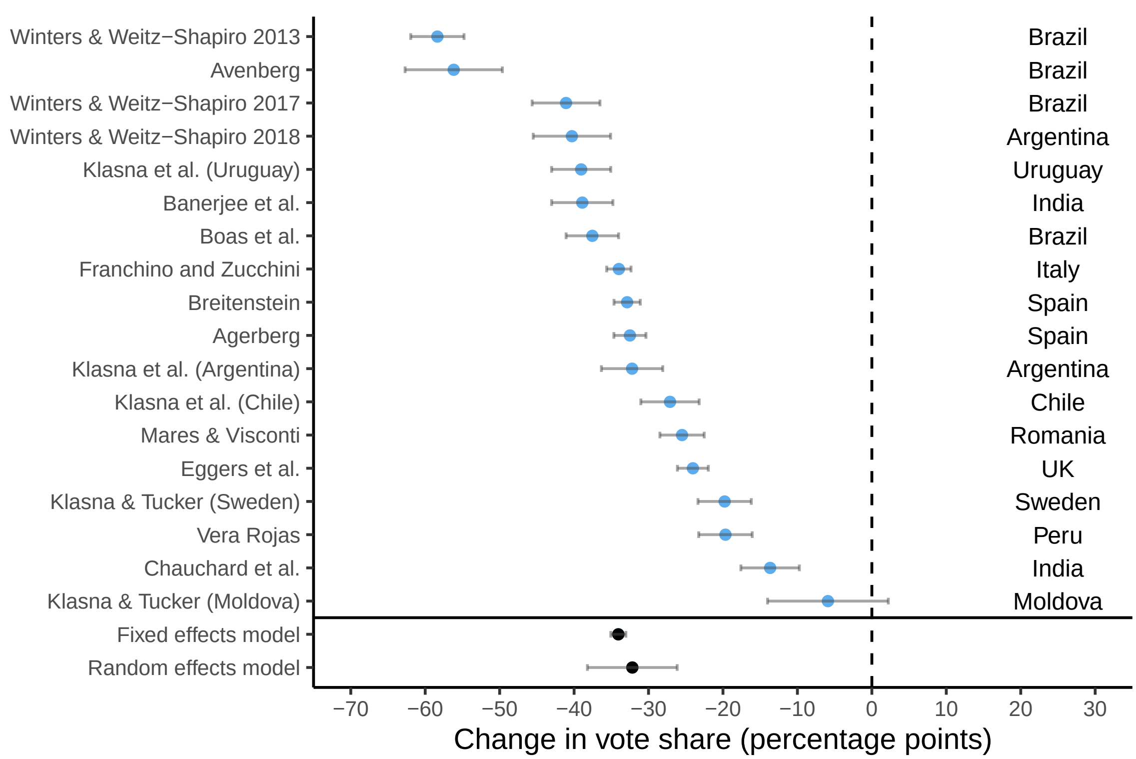

Figure 2: Survey experiments: Average treatment effect of corruption information on incumbent vote share and 95% confidence intervals

5 Exploring the discrepancy

What accounts for the large difference in treatment effects between field and survey experiments? One possibility is publication bias. Null results may be less likely to be published than significant results, particularly in a survey setting. A second possibility is social desirability bias, which may cause respondents in survey experiments to report normatively desirable outcomes. Alternatively, survey settings may not mirror real-world voting decisions. Potential ways in which the survey setting may differ from the field are: treatment salience and noncompliance, differences in outcome choices, and costliness/decision complexity. Weak treatments and noncompliance may decrease treatment effect sizes in field experiments. Design decisions may change the choice sets available to respondents. Finally, surveys may not capture the complexity and costliness of real-world voting decisions. It is possible that more complex factorial designs - such as conjoint experiments - may more successfully approximate real-world settings, and by extension field experiments. However, common methods of analysis of conjoint experiments may not capture all theoretical quantities of interest.

5.1 Publication bias and p-hacking

Of the ten field experiments I located, only six are published. By contrast, for survey experiments only three of the 16 papers remain unpublished, and these are recent drafts. This may reflect that the null results from field experiments are less likely to be published than their survey counterparts with large and highly significant negative treatment effects, even when standard errors are relatively small. While recognizing that the sample size of studies is small, OLS and logistic regression do not indicate that reported p-value is a significant predictor of publication status, although the directionality of coefficients is consistent with lower p-values being more likely to be published (Table A.9). However, this simple analysis is complicated by the fact that the p-value associated with the average treatment effect across all subjects may not be the primary p-value of interest in the paper.

In order to more formally test for publication bias, I first use the p-curve (Simonsohn, Simmons and Nelson 2015; Simonsohn, Nelson and Simmons 2014*a*,1). The p-curve is based on the premise that only "significant" results are typically published, and depicts the distribution of statistically significant p-values for a set of published studies. The shape of the p-curve is indicative of whether or not the results of a set of studies are derived from true effects, or from publication bias. If effect sizes are clustered around 0.05 (i.e. the p-curve is "left skewed"), this may be evidence of p-hacking, indicating that studies with p-values just below 0.05 are "selectively reported." If the p-curve is "right skewed" and there are more low p-values (0.01), this is evidence of true effects. All significant survey experimental results included in the meta-analysis are significant at the 1% level (making construction of a "curve" with bins of width 0.01 impossible), implying that publication bias likely does not explain the large negative treatment effects in survey experiments.[4] For field experiments, there is not a large enough number of published experiments to make the p-curve viable. Only six studies are published, and of these only four are significant at at least the 5% level.

Next, I test for publication bias by examining funnel plot asymmetry. A funnel plot depicts the outcomes from each study on the x-axis and their corresponding standard errors on the y-axis. The chart is overlaid with an inverted triangular confidence interval region (i.e. the "funnel"), which should contain 95% of the studies if there is no bias or between study heterogeneity. If studies with insignificant results remain unpublished the funnel plot may be asymmetric. Both visual inspection and regression tests of funnel plot asymmetry reveal an asymmetric funnel plot when survey and field experiments are grouped together (see Figure A.4 and Table A.10). However, this asymmetry disappears when accounting for heterogeneity by type of experiment, either with the inclusion of a survey experiment moderator (dummy) variable or by analyzing field and survey experiments separately (see Figure A.5, Figure A.6, Figure A.8, and Table A.10).

In sum, while publication bias cannot be ruled out completely—particularly with such a

[4]There is also no indication of publication bias at the 1% level using this method.

small sample size of field experiments—there is no smoking gun. This implies that differences in experimental design likely account for the difference in the magnitude of treatment effects in field versus survey experiments, rather than publication bias.

5.2 Social desirability bias

A second possible explanation is social desirability or sensitivity bias, in which survey respondents under-report socially undesirable behavior. A respondent may think a particular response will be perceived unfavorably by society as whole, by the researcher(s), or both, and underreport such behavior. In the case of corruption, respondents are likely to perceive corruption as harmful to society, the economy, and their own personal well-being. They may therefore be more likely to choose the socially desirable option (no corruption), particularly when observed by a researcher or afraid of response disclosure. However, a researcher is not the only social referent to whom an respondent may wish to give a socially desirable response. Respondents also may not wish to admit to themselves that they would vote for a corrupt candidate. Voting against corruption in the abstract may therefore reflect the respondents' actual preferences.

However, sensitivity bias is unlikely to account entirely for the difference in magnitude of treatment effects. A recent meta-analysis finds that sensitivity biases are typically smaller than 10 percentage points and are in some domains approximately zero (Blair, Coppock and Moor 2018). Notably, Blair, Coppock and Moor (2018) find that respondents underreport vote buying by 8 percentage points on average. As vote buying is often considered a form of corruption, the amount of sensitivity bias present in corruption survey experiments may be similar.

How might we overcome social desirability bias in survey experiments? One option is to eschew hypothetical candidates in favor of real candidates. Of course, for ethical reasons this limits researchers to having actual information regarding the corrupt actions of candidates. An alternative option is the use of list experiments. None of the survey experiments included here are list experiments. More complex factorial designs such as conjoint experiments have also been claimed to reduce social desirability bias (Hainmueller, Hopkins and Yamamoto 2014; Horiuchi, Markovich and Yamamoto 2018).

A related but distinct source of bias is "hypothetical bias" found in stated preference surveys in environmental economics, in which respondents report a willingness to pay that is larger than what they will actually pay using their own money as the costs are purely hypothetical (Loomis 2011). For corruption experiments, this would manifest as respondents reporting a willingness to punish corruption larger than in reality as the costs in terms of tradeoffs are purely hypothetical. There are few costs to selecting the socially desirable option in a hypothetical survey experiment. By contrast, the cost of changing one's actual vote (as in field experiments) may be higher. Voters might have pre-existing favorable opinions of real candidates, discount corruption information, or have strong material and/or ideological incentives to stick with their candidate. As the informational treatment will only have an effect on supporters of the corrupt candidate who must change their vote—opponents have already decided not to vote for the candidate—these costs are particularly high.

5.3 Do surveys mirror field settings and real-world voting decisions?

Even if subjects (voters), treatments (information), and outcome (vote choice) are similar, contextual differences between survey and field experiments may also offer fundamentally different choice sets to voters. These discrepancies between survey and field experimental designs, as well as between the designs of different survey experiments, may alter respondents' potential outcomes and thus capture different estimands. Some possible contextual differences are discussed below.

5.3.1 Treatment strength and noncompliance

Informational treatments may be weaker in field experiments in part because of their method of delivery. Survey treatments tend to be clear and authoritative, and often provide information on the challenger (clean or corrupt). By contrast, many of the informational treatments utilized in past information and accountability field experiments—e.g. fliers, newspaper articles, and text messages—provide relatively weak one-time treatments that may even contain information subjects are already aware of. In fact, the natural experiment conducted by Ferraz and Finan (2008)—which takes advantage of random municipal corruption audits conducted by the Brazilian government—may provide evidence of the effectiveness of stronger treatments. The results of the audits were disseminated naturally by newspapers and political campaigns, and their study provides the largest estimated treatment effect amongst real-world experiments. While not measuring specific vote choice, past experiments using face-to-face canvassing contact have also demonstrated relatively large effects on voter turnout (Gerber and Green 2000; Green and Gerber 2019; Green, Gerber and Nickerson 2003; Kalla and Broockman 2018), but these methods have not been used in any information and accountability field experiments to date.

Treatment effects in field experiments (fliers, newspapers, etc.) may also be weaker in part because they can be missed by segments of the treatment group. More formally, survey experiments do not have noncompliance by design and therefore the average treatment effect (ATE) is equal to the intent-to-treat (ITT) effect, whereas field experiments present ITT estimates as they are unable to identify which individuals in the treatment area actually received and internalized the informational treatment. Ideally, we would calculate the complier average causal effect (CACE)—the average treatment effect among the subset of respondents who comply with treatment—in field experiments, but we are unfortunately unable to observe compliance in any of the corruption experiments conducted to date.

A theoretical demonstration shows how noncompliance can drastically alter the ITT. The ITT is defined as $ITT = CACE \times \pi_c$ where π_C indicates the proportion of compliers in the treatment group. When $\pi_C = 1$, $ITT = CACE = ATE$. If the ITT = -0.0018—as fixed-effects meta-analysis estimates in field experiments—but only 10% of treated individuals "complied" with the treatment by reading the flier sent to them, this implies that the CACE is $\frac{-0.0018}{0.1} = -0.018$, or approximately -2 percentage points. In other words, while the effect of receiving a flier is roughly 0.2 percentage points, the effect of *reading* the flier is -2 percentage points. As the $ITT = \pi_c \times CACE$, any noncompliance necessarily reduces the size of the ITT. However, for the CACE to be equal in both survey and field experiments, the compliance rate in field experiments would have to be a miniscule 0.5%, implying that compliance likely does not tell the whole story. Future field experiments should therefore make an effort to build measurement of noncompliance into the research design.

Finally, treatments may be less salient at the time of vote choice in a field setting. Survey treatments are directly presented to respondents who are forced to immediately make a vote choice. Kalla and Broockman (2018) note that this mechanism manifests in campaign contact field experiments, where contact long before election day followed by immediate measurement of outcomes appears to persuade voters, whereas there is a null effect on vote choice on election day. Similarly, Sulitzeanu-Kenan, Dotan and Yair (2019) show that increasing the salience of corruption can increase electoral sanctioning, even without providing any new corruption information. Weaker treatments or lower salience of corruption in field experiments will weaken the treatment effect even amongst compliers (i.e. the CACE), further reducing the ITT.

Non-compliance, weak treatments, and declining treatment salience over time therefore make it unclear if the zero and null effects observed in field experiments stem from methodological choices or an actual lack of preference updating. Future field experiments should therefore consider using stronger treatments that simultaneously allow for measurement of noncompliance (e.g. canvassing).

5.3.2 Outcome choice

While vote choice is the outcome variable across all of the experiments investigated here, the choice set offered to voters is not necessarily always identical. Consider a voter's choice between two candidates in a field experiment conducted during an election. A candidate is revealed to be corrupt to voters in a treatment group, but not to voters in control. The treated voter can cast a ballot for corrupt candidate A, or candidate B, who may be clean or corrupt. The control voter can cast a ballot for candidate A or candidate B, and has no corruption information. Now consider a survey experiment with a vignette in which the randomized treatment is whether the corrupt actions of a politician are revealed or not. The treated voter can vote for the corrupt candidate A or not, but no challenger exists. Likewise, the control voter can vote for clean candidate A or not, but no challenger exists. Conjoint experiments overcome this difference, but the option to abstain still does not exist in the survey setting.[5] These differences in design offer fundamentally different choice sets to voters, altering respondents' potential outcomes and thus capturing different estimands.

5.3.3 Complexity, costliness, and conjoint experiments

Previous researchers have noted that even if voters generally find corruption distasteful, the quality of the information provided or positive candidate attributes and policies may outweigh the negative effects of corruption to voters, mitigating the effects of information provision on vote share. [6] These mitigating factors will naturally arise in a field setting, but may only be salient to respondents if specifically manipulated in a survey setting.

A number of survey experiments have therefore added factors other than corruption as mitigating variables, such as information quality, policy, economic benefit, and co-partisanship. Studies have randomized the quality of corruption information[7] (Banerjee et al. 2014; Botero et al. 2015; Breitenstein 2019; Mares and Visconti 2019; Weitz-Shapiro and Winters 2017; Winters and Weitz-Shapiro 2018), finding that lower quality information produces smaller negative treatment effects (see Figure A.9). Policy stances in line with voter preferences have also been shown to mitigate the impact of corruption (Franchino and Zucchini 2015; Rundquist, Strom and Peters 1977). Evidence also suggests that respondents are more forgiving of corruption when it benefits them economically (Klašnja, Lupu and Tucker 2017; Winters and Weitz-Shapiro 2013). Evidence of co-partisanship as a limiting factor to cor-

[5] See Eggers, Vivyan and Wagner (2018) and Agerberg (2019) for exceptions.

[6] See De Vries and Solaz (2017) for a comprehensive overview.

[7] For example, accusations from an independent anti-corruption authority may be deemed more credible than those from an opposition party, and accusations may be deemed less credible than a conviction.

ruption deterrence is mixed. Anduiza, Gallego and Muñoz (2013), Agerberg (2019), and Breitenstein (2019) show that co-partisanship decreases the importance of corruption to Spanish respondents in survey experiments. Solaz, De Vries and de Geus (2018) induce in-group attachment in a lab-experiment of UK subjects, finding that in-group membership reduces sanction of "corrupt" participants. However, Klašnja, Lupu and Tucker (2017) find relatively small effects of co-partisanship in Argentina, Chile, and Uruguay, and Rundquist, Strom and Peters (1977) find null effects in the US in the 1970s. Konstantinidis and Xezonakis (2013) also find that partisanship does not moderate punishment of corruption in a survey experiment in Greece. Boas, Hidalgo and Melo (2018) posit that abandoning dynastic candidates is particularly costly in Brazil. This evidence suggests that voters punish corruption less when it is costly to do so, and that these costly factors may differ by country.

The fact that moderating variables may dampen the salience of corruption to voters has clearly not been lost on previous researchers. However, in the field setting numerous moderating factors may be salient to the voter. Conjoint experiments allow researchers to randomize many candidate characteristics simultaneously, and thus have become a popular survey method for investigating the relative weights respondents give to different candidate attributes. In addition, conjoints force respondents to pick between two candidates, better emulating the choice required in an election. Finally, conjoints may minimize social desirability bias as they reduce the probability that the respondent is aware of the researcher's primary experimental manipulation of interest (e.g. corruption).[8]

However, researchers have thus far tended to present the results of conjoint experiments as average marginal component effects (AMCEs), then compare the magnitude of these effect sizes. AMCEs represents the unconditional marginal effect of an attribute (e.g. corruption) averaged over all possible values of the other attributes. This measurement is often ex-

[8]This is explicitly mentioned by Hainmueller, Hopkins and Yamamoto (2014), who argue that conjoint experiments give respondents "various attributes and thus [they] can often find multiple justifications for a given choice." Note, however, that an experiment does not necessarily need to be a conjoint design to have this feature. Conjoint experiments encourage researchers to randomize more attributes and therefore typically contain more complex hypothetical vignettes. However, the same vignette complexity could be achieved without full randomization of these attributes.

tremely valuable, and crucially allows researchers to test multiple causal hypotheses and compare relative magnitudes of effects between treatments. However, this may or may not be a measure of substantive interest to the researcher, and in fact implies that the AMCE is dependent on the distribution of the other attributes in the experiment.[9] These attributes are usually completely randomized and therefore uniform. However, in the real world, candidate attributes are not fully randomized and uniformly distributed, so external validity is questionable. When we have a primary treatment of interest, such as corruption, we therefore want to see how a "typical candidate" is punished for corruption, but a typical candidate is not a randomized candidate, he or she is a candidate designed to appeal to voters—even if they are revealed to be corrupt. The corruption AMCE is therefore valid in the context of the experiment—marginalizing over the distribution of all of the other attributes in the experiment—but would likely be much smaller for a realistic candidate. This perhaps implies that AMCEs are best utilized when the experiment contains the entire universe of possible attributes—for example to find an optimal policy design from the perspective of voter preference.

When researchers have strong theories about the conditions that shape voter decision-making, a more appropriate method may be to calculate average marginal effects in order to present predicted probabilities of voting for a candidate under these conditions.[10] For example, in a conjoint experiment including corruption information, this might be interpreted as the probability of voting for a candidate that is both corrupt and possesses other particular feature levels (e.g. party membership and/or policy positions), marginalizing across all other features in the experiment.

To illustrate this point, I replicate the conjoint experiments conducted in Spain by Breitenstein (2019) and in Italy by Franchino and Zucchini (2015), and present both AMCEs and

[9]See de la Cuesta, Egami and Imai (2019) for additional discussion and empirical demonstration of the impact of choice of distribution on the AMCE.

[10]This method is utilized by Teele, Kalla and Rosenbluth (2018) to examine the probability of voting for female or male candidates holding other candidate attributes (marital status and number of children) constant, and in corruption experiments by Breitenstein (2019) and Agerberg (2019). This method is discussed in more detail by Leeper, Hobolt and Tilley (2019).

predicted probabilities. The Breitenstein (2019) re-analysis is presented in the main text, while the re-analysis of Franchino and Zucchini (2015) is in the appendix. Note that I group all corruption accusation levels into a single "corrupt" level in my replications. The Breitenstein (2019) predicted probabilities are presented as a function of corruption, co-partisanship, political experience, and economic performance. The charts therefore show the probability of preferring a candidate who is always corrupt, but is a co-partisan or not, has low or high experience, and whose district experienced good or bad economic performance, marginalizing across all other features in the experiment. For Franchino and Zucchini (2015), the predicted probabilities are presented as a function of corruption and two policy positions - tax policy and same sex marriage - separately for conservative and liberal respondents. The charts therefore show the probability of preferring a candidate who is corrupt, but has particular levels of tax policy, and same sex marriage policy, marginalizing across all other features in the experiment. Note that Franchino and Zucchini (2015) correctly conclude that their typical "respondent prefers a corrupt but socially and economically progressive candidate to a clean but conservative one." While I therefore illustrate how predicted probabilities can be used to draw conclusions that may be masked by examination of AMCEs alone, the authors themselves do not make this mistake. I perform the same analysis including only cases where the challenger is clean in the appendix.

A casual interpretation of the traditional AMCE plots presented in Figure 3 and Figure A.12 suggests that it is very unlikely a corrupt candidate would be chosen by a respondent. By contrast, the predicted probabilities plots presented in Figure 4, Figure A.13, and Figure A.14 show that even for corrupt candidates in the conjoint, the right candidate or policy platform presented to the right respondents can garner over 50% of the predicted hypothetical vote. Further, the attributes included in these conjoints surely do not represent all candidate attributes relevant to voters, and indeed differ greatly across experiments. As in Agerberg (2019), the level of support for corrupt candidates also varies based on whether or not the challenger is clean (Figure A.10, Figure A.15, Figure A.14). In other words, respondents find it costly to abandon their preferences even if it forces them to select a corrupt candidate, and this costliness varies highly depending on contextual changes and choice of other attributes included in the experiments.

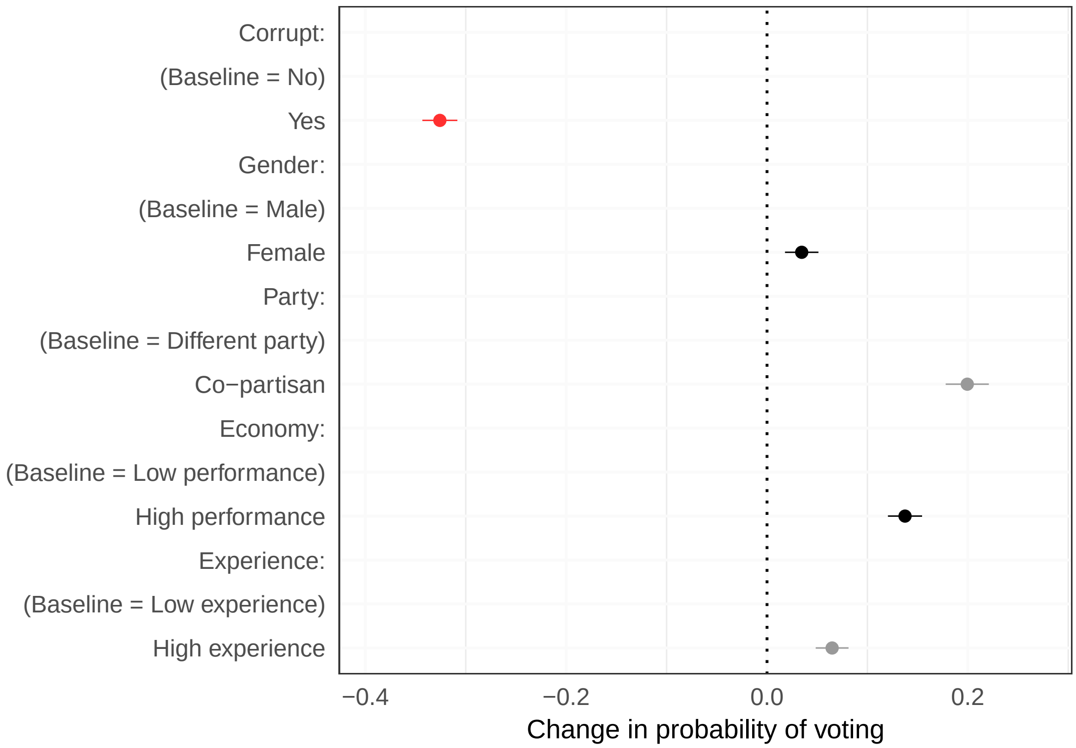

Figure 3: Breitenstein (2019) conjoint: average marginal component effects

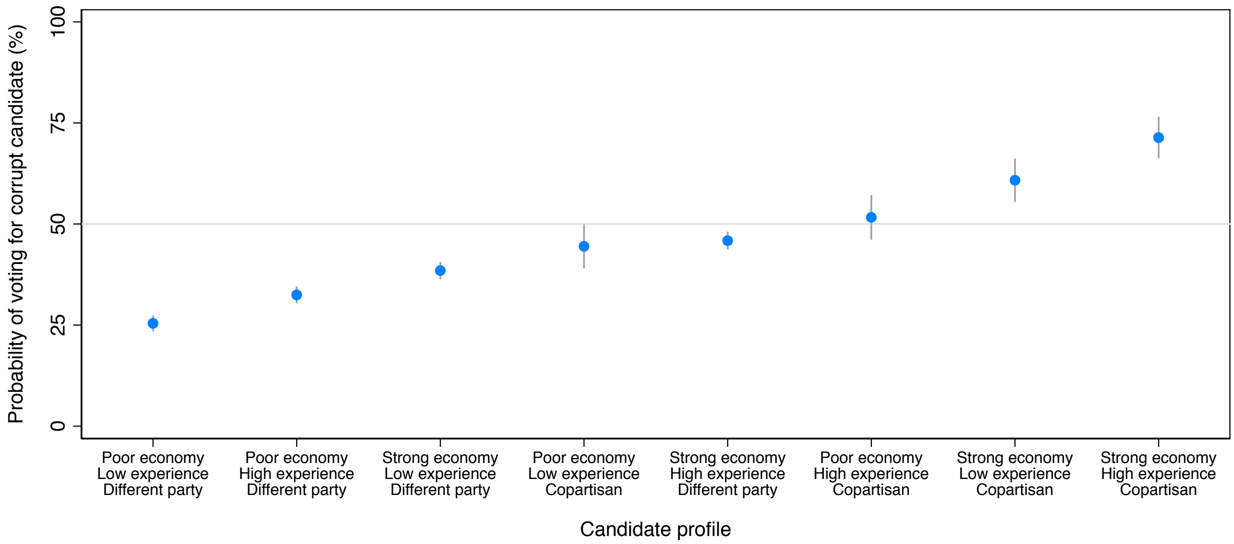

Figure 4: Breitenstein (2019) conjoint: can the right candidate overcome corruption?

Candidate or policy profiles that result in over 50% of voters selecting a corrupt candidate may not be outliers in real-world scenarios. Unlike in conjoint experiments, real-world candidates' attributes and policy profiles are not selected randomly, but rather represent choices designed to appeal to voters. Voters may also be unsure if the challenger is also corrupt or clean. It may therefore be preferable to analyze conjoint experiments as above, comparing outlier characteristics (e.g. corruption) to realistic candidate profiles that target specific voters, rather than fully randomized candidate profiles.

When the most theoretically relevant tradeoffs are unclear, we may be able to illuminate voter decision making processes through the use of decision trees. The decision tree in Figure 5 was trained using all randomized variables in the Breitenstein (2019) conjoint, and the tree was pruned to minimize cross-validated classification error rate. Figure 5 draws similar conclusions as the predicted probabilities chart shown in Figure 4 with respect to what factors matter most to voters. A similar figure depicting corrupt candidates facing clean challengers only can be found in Figure A.11.

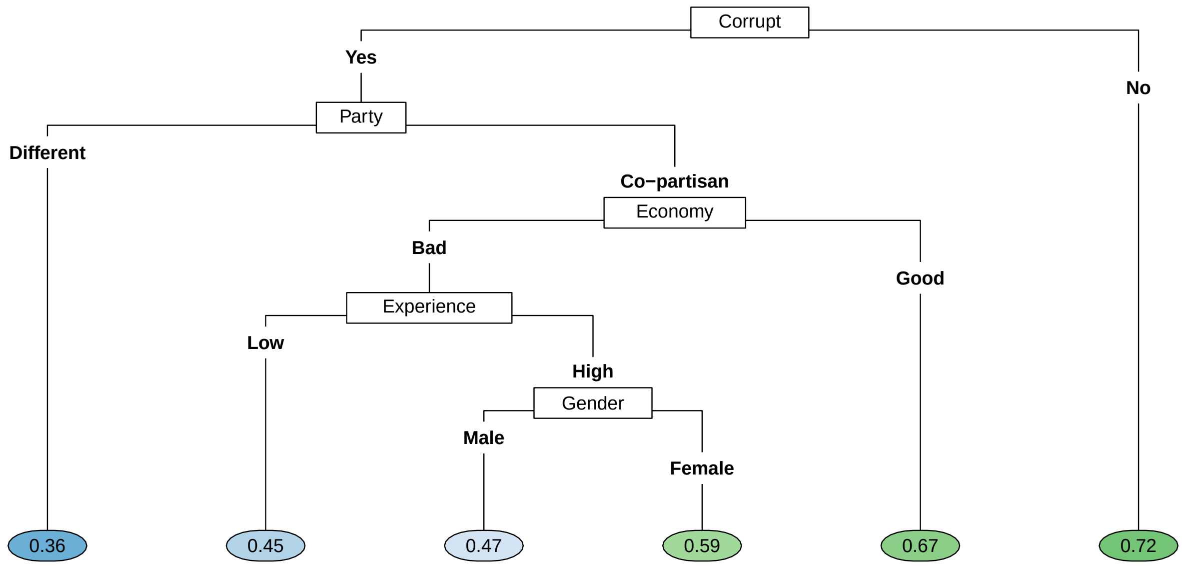

Figure 5: Breitenstein (2019) conjoint decision tree: predicted probabilities of voting for candidate

6 Discussion

The field experimental results reported here align with a growing body of literature that shows minimal effects of information provision on voting outcomes. The primary conclusion of the Metaketa I project—which sought to determine if politicians were rewarded for positive information and punished for negative information—was that "the overall effect of information [provision] across all studies is quite precisely estimated—and not statistically distinguishable from zero" (Dunning, Grossman, Humphreys, Hyde, McIntosh and Nellis 2018), and a meta-analysis by Kalla and Broockman (2018) suggests that the effect of campaign contact and advertising on voting outcomes in the United States is close to zero in general elections.

However, we should be careful not to conclude that voters never punish politicians for malfeasance from these experiments, or that field experiments necessarily provide the "real" estimates. Field and natural experiments in other domains have found effects when identifying persuadable voters prior to treatment delivery (Kalla and Broockman 2018; Rogers and Nickerson 2013), or when using higher dosage treatments (Adida, Gottlieb, Kramon and McClendon 2019; Ferraz and Finan 2008). Combining stronger treatments, measurement of non-compliance, and pre-identification of subgroups most susceptible to persuasion should therefore be a goal of future field experiments.

Many of the survey experimental studies discuss how their findings may partially stem from the particular conditions of the experiment, claim that they are only attempting to identify tradeoffs or mitigating effects, and/or acknowledge the limitations of external validity. However, other studies do not (see **??**). A common approach is to cite Hainmueller, Hangartner and Yamamoto (2015), who show similar effects in a vignette, conjoint, and natural experiment. However, Hainmueller, Hangartner and Yamamoto (2015) use closeness in the magnitude of treatment effects between vignettes and the natural experiment as a justification for correspondence between the two methodologies. Their study therefore suggests that the relative importance *and magnitude* of treatment effects should be similar between hypothetical vignettes and the real world, which this meta-analysis shows is not the case with corruption voting. Further, the natural experimental benchmark takes the form of a survey/leaflet sent to voters containing the attributes of immigrants applying for naturalization in Swiss municipalities. The conjoint experiment is therefore able to perfectly mimic the amount of information voters posses in the real world, which is not the case for political candidates.[11] We should therefore be cautious when extrapolating the correspondence between these studies to additional cases.

7 Conclusion

In an effort to test whether voters adequately hold politicians accountable for malfeasance, researchers have turned to experimental methods to measure the causal effect of learning

[11] Hainmueller, Hangartner and Yamamoto (2015) acknowledge this directly, stating that "these data provide an ideal behavioral benchmark to evaluate stated preference experiments, because they closely resemble a real-world vignette experiment" and that "unlike many other real-world choice situations, in the referendums, the information environment and choice attributes are sufficiently constrained, such that they can be accurately mimicked in a survey experimental design."

about politician corruption on vote choice. A meta-analytic assessment of these experiments reveals that conclusions differ drastically depending on whether the experiment was deployed in the field and monitored actual vote choice, versus hypothetical vote choice in an online setting. Across field experiments, the aggregate treatment effect of providing information about corruption on vote share is approximately zero. By contrast, in survey experiments corrupt candidates are punished by respondents by approximately 32-34 percentage points.

I explore publication bias, social desirability bias, and contextual differences in the nature of the experimental designs as possible explanations for the discrepancy between field and survey experimental results. I do not find systematic evidence of publication bias. Social desirability bias may drive some of the difference if survey experiments capture anti-corruption norms rather than realistic voter behavior. The survey setting may differ from the field due to contextual differences such as noncompliance, treatment strength, differences in outcome choice sets, and costliness/decision complexity. Noncompliance necessarily decreases treatment effect sizes in field experiments. Less clear treatments or lower salience of information to voters on election day versus immediately after treatment receipt will also reduce effect sizes. Previous survey experiments have also shown that treatment effects diminish as the costliness of changing one's vote increases, and these costs are likely to be much higher and more multitudinous in an actual election. The personal cost of changing one's vote may therefore be higher than accepting corruption in many real elections, but not in surveys.

High-dimension factorial designs such as conjoint experiments may better capture the costly tradeoffs voters make in the survey setting. However, it may be preferable to analyze vote-choice conjoint experiments by comparing the probability of voting for a realistic candidate with outlier characteristics (e.g. corruption) to the probability of voting for the same realistic candidate without this characteristic, rather than examining differences in AMCEs across fully randomized candidate profiles.

These findings suggest that while vote-choice survey experiments may provide information on the directionality of informational treatments in hypothetical scenarios, the point estimates they provide may not be representative of real-world voting behavior. More generally, researchers should exercise caution when interpreting actions taken in hypothetical vignettes as indicative of real world behavior such as voting.

References

Adida, Claire, Jessica Gottlieb, Eric Kramon and Gwyneth McClendon. 2019. "Under what conditions does performance information influence voting behavior? Lessons from Benin." *Metaketa I: The Limits of Electoral Accountability* .

Agerberg, Mattias. 2019. "The Lesser Evil? Corruption Voting and the Importance of Clean Alternatives." *Comparative Political Studies* p. 0010414019852697.

Anduiza, Eva, Aina Gallego and Jordi Muñoz. 2013. "Turning a blind eye: Experimental evidence of partisan bias in attitudes toward corruption." *Comparative Political Studies* 46(12):1664–1692.

Arias, Eric, Horacio Larreguy, John Marshall and Pablo Querubin. 2018. Priors Rule: When do Malfeasance Revelations Help or Hurt Incumbent Parties? Technical report National Bureau of Economic Research.

URL: *https://www.nber.org/papers/w24888*

Arias, Eric, Horacio Larreguy, John Marshall and Pablo Querubin. 2019. "Information Provision, Voter Coordination, and Electoral Accountability: Evidence from Mexican Social Networks." *American Political Science Review* .

Arvate, Paulo and Sergio Mittlaender. 2017. "Condemning corruption while condoning inefficiency: an experimental investigation into voting behavior." *Public Choice* 172(3-4):399–419.

Avenburg, Alejandro. 2019. "Public Costs versus Private Gain: Assessing the Effect of Different Types of Information about Corruption Incidents on Electoral Accountability." *Journal of Politics in Latin America* 11(1):71–108.

Azfar, Omar and William Robert Nelson. 2007. "Transparency, wages, and the separation of powers: An experimental analysis of corruption." *Public Choice* 130(3-4):471–493.

Banerjee, Abhijit, Donald Green, Jennifer Green and Rohini Pande. 2010. Can voters be primed to choose better legislators? Experimental evidence from rural India. In

Presented and the Political Economics Seminar, Stanford University. Citeseer.

URL: *https://pdfs.semanticscholar.org/a204/eb3e92d382dd312790f47df9aefde657fd13.pdf*

Banerjee, Abhijit, Donald P Green, Jeffery McManus and Rohini Pande. 2014. "Are poor voters indifferent to whether elected leaders are criminal or corrupt? A vignette experiment in rural India." *Political Communication* 31(3):391–407.

Banerjee, Abhijit, Selvan Kumar, Rohini Pande and Felix Su. 2011. "Do informed voters make better choices? Experimental evidence from urban India." *Unpublished manuscript* .

URL: *https://pdfs.semanticscholar.org/45aa/1e275e770103f7a7d7b02ba86fb46afa89c0.pdf*

Blair, Graeme, Alexander Coppock and Margaret Moor. 2018. When to Worry About Sensitivity Bias: Evidence from 30 Years of List Experiments. Technical report Working Paper.

URL: *https://alexandercoppock.com/papers/BCM_list.pdf*

Boas, Taylor C, F Daniel Hidalgo and Marcus André Melo. 2018. "Norms versus Action: Why Voters Fail to Sanction Malfeasance in Brazil." *American Journal of Political Science* .

Botero, Sandra, Rodrigo Castro Cornejo, Laura Gamboa, Nara Pavao and David W Nickerson. 2015. "Says who? An experiment on allegations of corruption and credibility of sources." *Political Research Quarterly* 68(3):493–504.

Breitenstein, Sofia. 2019. "Choosing the crook: A conjoint experiment on voting for corrupt politicians." *Research & Politics* 6(1):2053168019832230.

Buntaine, Mark T, Ryan Jablonski, Daniel L Nielson and Paula M Pickering. 2018. "SMS texts on corruption help Ugandan voters hold elected councillors accountable at the polls." *Proceedings of the National Academy of Sciences* 115(26):6668–6673.

Camerer, Colin. 2011. "The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List." *Available at SSRN 1977749* .

Chauchard, Simon, Marko Klašnja and SP Harish. 2019. "Getting Rich Too Fast? Voters? Reactions to Politicians? Wealth Accumulation." *The Journal of Politics* 81(4):1197–1209.

Chong, Alberto, Ana L De La O, Dean Karlan and Leonard Wantchekon. 2014. "Does corruption information inspire the fight or quash the hope? A field experiment in Mexico on voter turnout, choice, and party identification." *The Journal of Politics* 77(1):55–71.

Coppock, Alexander and Donald P Green. 2015. "Assessing the correspondence between experimental results obtained in the lab and field: A review of recent social science research." *Political Science Research and Methods* 3(1):113–131.

De Figueiredo, Miguel FP, F Daniel Hidalgo and Yuri Kasahara. 2011. "When do voters punish corrupt politicians? Experimental evidence from Brazil." *Unpublished manuscript, UC Berkeley* .

URL: https://law.utexas.edu/wp-content/uploads/sites/25/figueiredo_when_do_voters_punish.pdf

de la Cuesta, Brandon, Naoki Egami and Kosuke Imai. 2019. "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution." *Working Paper* .

De Vries, Catherine E and Hector Solaz. 2017. "The electoral consequences of corruption." *Annual Review of Political Science* 20:391–408.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D Hyde, Craig McIntosh and Gareth Nellis. 2018. "Metaketa I: Information, accountability, and cumulative learning.".

Eggers, Andrew C, Nick Vivyan and Markus Wagner. 2018. "Corruption, accountability, and gender: do female politicians face higher standards in public life?" *The Journal of Politics* 80(1):321–326.

Ferraz, Claudio and Frederico Finan. 2008. "Exposing corrupt politicians: the effects of Brazil's publicly released audits on electoral outcomes." *The Quarterly Journal of Economics* 123(2):703–745.

Franchino, Fabio and Francesco Zucchini. 2015. "Voting in a multi-dimensional space: a conjoint analysis employing valence and ideology attributes of candidates." *Political Science Research and Methods* 3(2):221–241.

Gerber, Alan S and Donald P Green. 2000. "The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment." *American political science review* 94(3):653–663.

Gray, Cheryl W and Daniel Kaufman. 1998. "Corruption and development.".

Green, Donald P, Adam Zelizer, David Kirby et al. 2018. "Publicizing Scandal: Results from Five Field Experiments." *Quarterly Journal of Political Science* 13(3):237–261.

Green, Donald P and Alan S Gerber. 2019. *Get out the vote: How to increase voter turnout.* Brookings Institution Press.

Green, Donald P, Alan S Gerber and David W Nickerson. 2003. "Getting out the vote in local elections: Results from six door-to-door canvassing experiments." *Journal of Politics* 65(4):1083–1096.

Hainmueller, Jens, Daniel J Hopkins and Teppei Yamamoto. 2014. "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments." *Political Analysis* 22(1):1–30.

Hainmueller, Jens, Dominik Hangartner and Teppei Yamamoto. 2015. "Validating vignette and conjoint survey experiments against real-world behavior." *Proceedings of the National Academy of Sciences* 112(8):2395–2400.

Horiuchi, Yusaku, Zachary D Markovich and Teppei Yamamoto. 2018. "Can Conjoint Analysis Mitigate Social Desirability Bias?".

Kalla, Joshua L and David E Broockman. 2018. "The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments." *American Political*

Science Review 112(1):148–166.

Klašnja, Marko and Joshua A Tucker. 2013. "The economy, corruption, and the vote: Evidence from experiments in Sweden and Moldova." *Electoral Studies* 32(3):536–543.

Klašnja, Marko, Noam Lupu and Joshua A Tucker. 2017. "When Do Voters Sanction Corrupt Politicians?".

Kolstad, Ivar and Arne Wiig. 2009. "Is transparency the key to reducing corruption in resource-rich countries?" *World development* 37(3):521–532.

Konstantinidis, Iannis and Georgios Xezonakis. 2013. "Sources of tolerance towards corrupted politicians in Greece: The role of trade offs and individual benefits." *Crime, Law and Social Change* 60(5):549–563.

Leeper, Thomas J, Sara B Hobolt and James Tilley. 2019. "Measuring Subgroup Preferences in Conjoint Experiments." *Political Analysis* .

Loomis, John. 2011. "What's to know about hypothetical bias in stated preference valuation studies?" *Journal of Economic Surveys* 25(2):363–370.

Mares, Isabela and Giancarlo Visconti. 2019. "Voting for the lesser evil: Evidence from a conjoint experiment in Romania." *Political Science Research and Methods* .

Muñoz, Jordi, Eva Anduiza and Aina Gallego. 2012. Why do voters forgive corrupt politicians? Cynicism, noise and implicit exchange. In *International Political Science Association Conference, Madrid, Spain*.

Rogers, Todd and David Nickerson. 2013. "Can Inaccurate Beliefs About Incumbents be Changed? And Can Reframing Change Votes?".

Rose-Ackerman, Susan and Bonnie J Palifka. 2016. *Corruption and government: Causes, consequences, and reform*. Cambridge university press.

Rundquist, Barry S, Gerald S Strom and John G Peters. 1977. "Corrupt politicians and their electoral support: some experimental observations." *American Political Science Review* 71(3):954–963.

Simonsohn, Uri, Joseph P Simmons and Leif D Nelson. 2015. "Better P-curves: Making

P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015).”.

Simonsohn, Uri, Leif D Nelson and Joseph P Simmons. 2014*a*. “P-curve: a key to the file-drawer.” *Journal of Experimental Psychology: General* 143(2):534.

Simonsohn, Uri, Leif D Nelson and Joseph P Simmons. 2014*b*. “p-curve and effect size: Correcting for publication bias using only significant results.” *Perspectives on Psychological Science* 9(6):666–681.

Solaz, Hector, Catherine E De Vries and Roosmarijn A de Geus. 2018. “In-Group Loyalty and the Punishment of Corruption.” *Comparative Political Studies* p. 0010414018797951.

Stokes, Susan C, Thad Dunning, Marcelo Nazareno and Valeria Brusco. 2013. *Brokers, voters, and clientelism: The puzzle of distributive politics*. Cambridge University Press.

Sulitzeanu-Kenan, Raanan, Yoav Dotan and Omer Yair. 2019. “Can Institutions Make Voters Care about Corruption?” *The Journal of Politics* .

Teele, Dawn Langan, Joshua Kalla and Frances Rosenbluth. 2018. “The Ties That Double Bind: Social Roles and Women’s Underrepresentation in Politics.” *American Political Science Review* 112(3):525–541.

Vera, Sofia B. 2019. “Accepting or Resisting? Citizen Responses to Corruption Across Varying Levels of Competence and Corruption Prevalence.” *Political Studies* p. 0032321719868210.

Weitz-Shapiro, Rebecca and Matthew S Winters. 2017. “Can citizens discern? Information credibility, political sophistication, and the punishment of corruption in Brazil.” *The Journal of Politics* 79(1):60–74.

Weschle, Simon. 2016. “Punishing personal and electoral corruption: Experimental evidence from India.” *Research & Politics* 3(2):2053168016645136.

Winters, Matthew S and Rebecca Weitz-Shapiro. 2013. “Lacking information or condoning corruption: When do voters support corrupt politicians?” *Comparative Politics* 45(4):418–436.

Winters, Matthew S and Rebecca Weitz-Shapiro. 2015. "Political corruption and partisan engagement: evidence from Brazil." *Journal of Politics in Latin America* 7(1):45–81.

Winters, Matthew S and Rebecca Weitz-Shapiro. 2016. "Who?s in charge here? Direct and indirect accusations and voter punishment of corruption." *Political Research Quarterly* 69(2):207–219.

Winters, Matthew S and Rebecca Weitz-Shapiro. 2018. "Information credibility and responses to corruption: a replication and extension in Argentina." *Political Science Research and Methods* pp. 1–9.

A Appendix

A.1 *Lab experiments*

Table A.1: Lab experiments

Study	Country	ATE
Arvate and Mittlaender (2017)	Brazil	Negative
Azfar and Nelson (2007)	USA	Negative
Rundquist, Strom and Peters (1977)[1]	USA	Negative
Solaz, De Vries and de Geus (2018)	UK	Negative

[1] The candidate is always corrupt in the Rundquist, Strom and Peters (1977) experiment. A "corruption" point estimate is therefore not provided in the coefficient plot below.

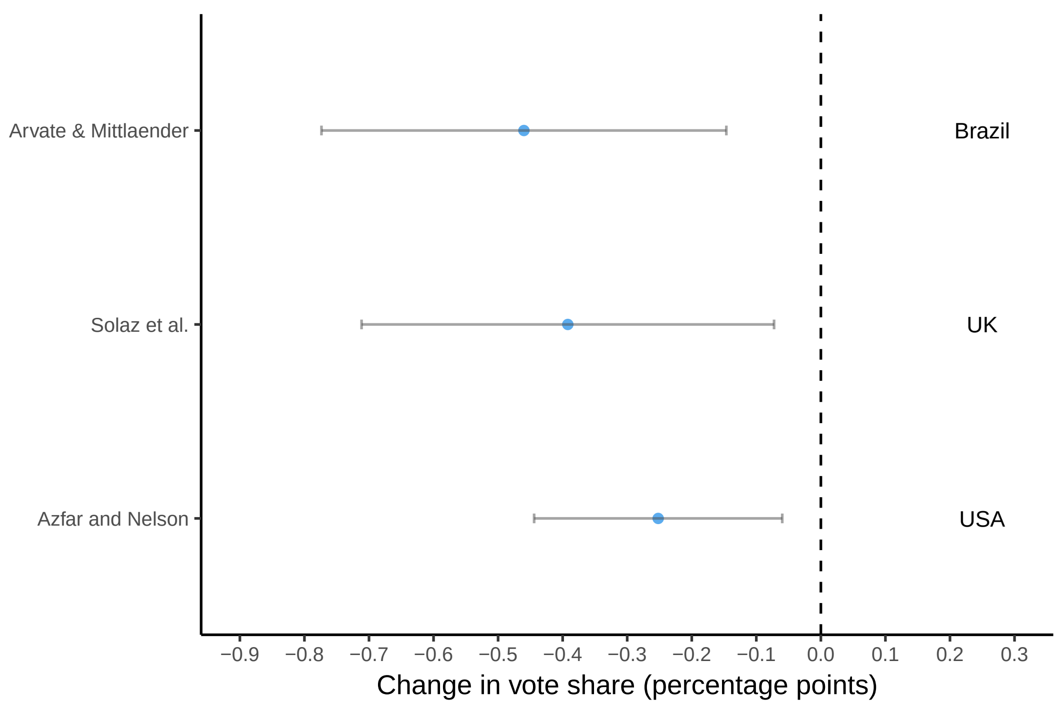

Figure A.1: Lab experiments: Average treatment effect of corruption information on vote share

.

A.2 Excluded studies

Table A.2: Excluded experiments

Study	Type	Reason for exclusion
Anduiza, Gallego and Muñoz (2013)	Survey	Lack of no-corruption control group
Botero et al. (2015)	Survey	Lack of no-corruption control group
De Figueiredo, Hidalgo and Kasahara (2011)	Survey	Outcome is hypothetically changing actual vote
Green, Zelizer, Kirby et al. (2018)	Field	Outcome is favorability rating, not vote share
Konstantinidis and Xezonakis (2013)	Survey	Lack of no-corruption control group
Muñoz, Anduiza and Gallego (2012)	Survey	Lack of no-corruption control group
Rundquist, Strom and Peters (1977)	Lab	Lack of no-corruption control group
Weitz-Shapiro and Winters (2017)	Survey	Data identical to Winters and Weitz-Shapiro (2016)
Winters and Weitz-Shapiro (2015)	Survey	Data identical to Winters and Weitz-Shapiro (2013)
Weschle (2016)	Survey	Lack of no-corruption control group

A.3 Meta-analysis and heterogeneity by type of experiment

Table A.3: Meta-analysis by type of experiment

Value	Estimate
Field: weighted fixed effects	-0.002
	(0.014)
Field: random effects	-0.001
	(0.008)
Survey: weighted fixed effects	-0.341
	(0.005)
Survey: random effects	-0.322
	(0.005)

Note: Standard errors in parenthesis. Figures rounded to nearest thousandth decimal place.

Table A.4: Random effects meta-analysis (all studies)

Value	Estimate
Estimate	-0.209
	(0.035)
Estimated total heterogeneity	0.034
	(0.01)

Note: Standard errors in parenthesis. Figures rounded to nearest thousandth decimal place.

Table A.5: Mixed effects meta-analysis with survey experiment moderator

Value	Estimate
Constant	-0.005
	(0.034)
Survey experiment moderator	-0.317
	(0.042)
Residual heterogenity with moderator	0.011
	(0.003)
Heterogenity accounted for	68.306%

Note: Standard errors in parenthesis. Figures rounded to nearest thousandth decimal place.

"Heterogeneity accounted for" is calculated as: $\frac{\text{(Total heterogeneity} - \text{Residual heterogeneity)}}{\text{(Total heterogeneity)}}$

A.4 Robustness checks

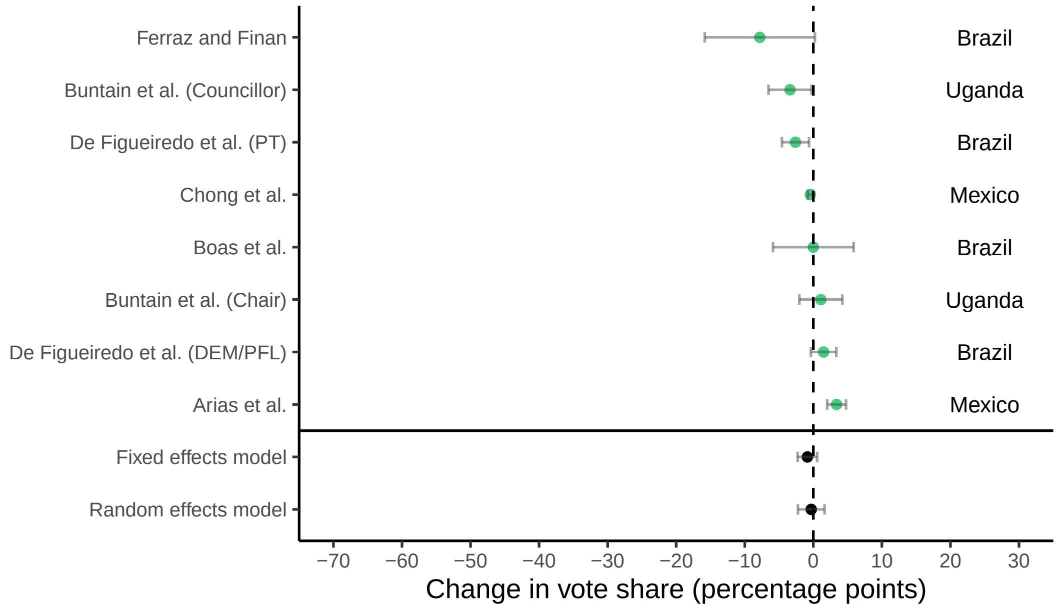

Figure A.2: Field experiments: Average treatment effect of corruption information on incumbent vote share (excluding Banerjee et al. (2010) and Banerjee et al. (2011))

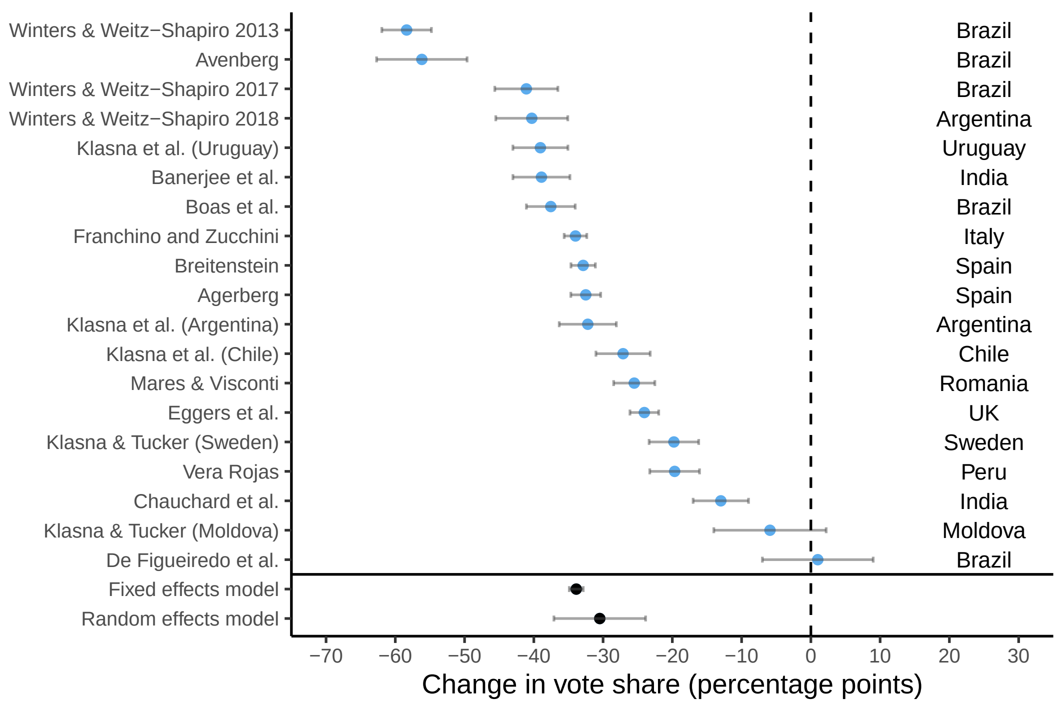

Figure A.3: Survey experiments: Average treatment effect of corruption information on incumbent vote share (including De Figueiredo, Hidalgo and Kasahara (2011))

.

Table A.6: Meta-analysis (all field experiments excluding Banerjee et al. (2010) and Banerjee et al. (2011))

Value	Estimate
Field: weighted fixed effects	-0.009
	(0.007)
Field: random effects	-0.003
	(0.01)

Note: Standard errors in parenthesis. Figures rounded to nearest thousandth decimal place.

Table A.7: Random effects meta-analysis (all studies excluding Banerjee et al. (2010) and Banerjee et al. (2011))

Value	Estimate
Estimate	-0.225
	(0.036)
Estimated total heterogeneity	0.033
	(0.01)

Note: Standard errors in parenthesis. Figures rounded to nearest thousandth decimal place.

Table A.8: Mixed effects meta-analysis with survey experiment moderator (excluding Banerjee et al. (2010) and Banerjee et al. (2011))

Value	Estimate
Constant	-0.009
	(0.039)
Survey experiment moderator	-0.312
	(0.047)
Residual heterogenity with moderator	0.012
	(0.004)
Heterogenity accounted for	64.43%

Note: Standard errors in parenthesis. Figures rounded to nearest thousandth decimal place.

A.5 Publication bias

Table A.9: Do p-values predict publication status?

	Dependent variable:	
	Published	
	OLS	Logit
Reference: P less than 0.01	0.84***	1.67***
	(0.10)	(0.63)
P less than 0.05	−0.18	−0.98
	(0.27)	(1.38)
P less than 0.1	0.16	14.89
	(0.44)	(2,399.54)
P greater than 0.1	−0.44*	−2.08*
	(0.22)	(1.11)
Observations	28	28
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table A.10: Regression tests for funnel plot asymmetry

Studies included	p value
All	0.0003
All with moderator	0.932
Field	0.840
Survey	0.814

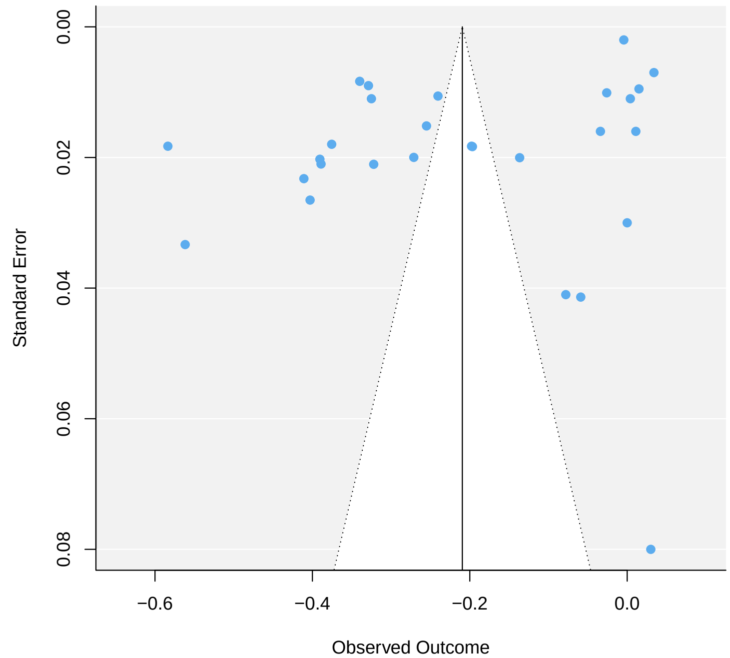

Figure A.4: Funnel plot: all experiments

.

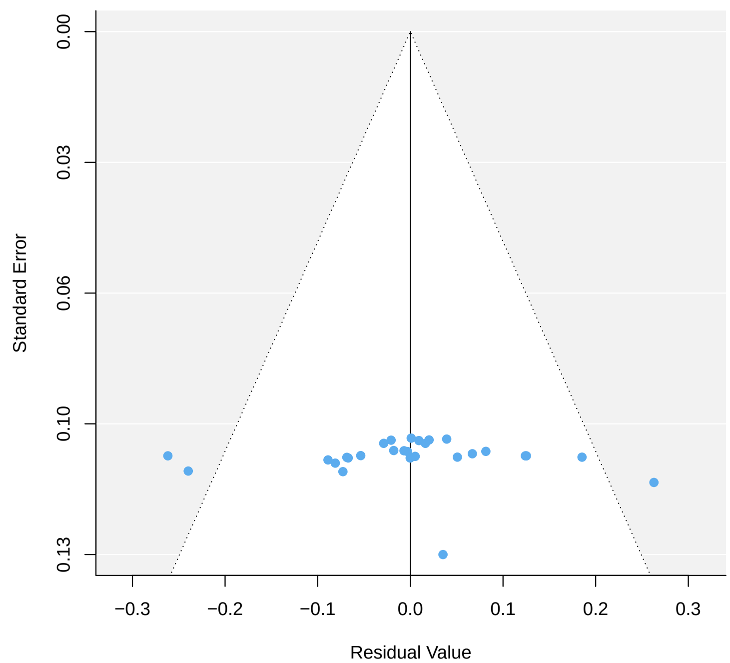

Figure A.5: Funnel plot: all experiments with field experiment moderator

.

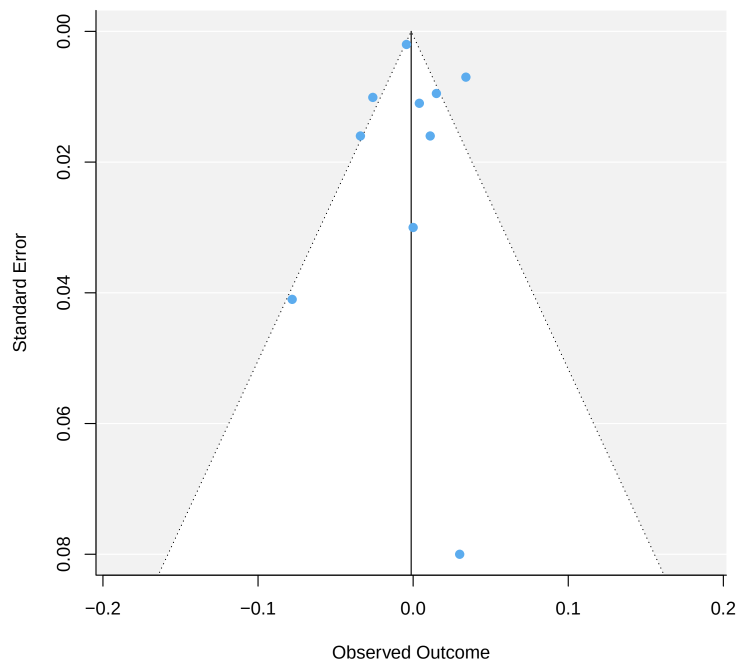

Figure A.6: Funnel plot: field experiments

.

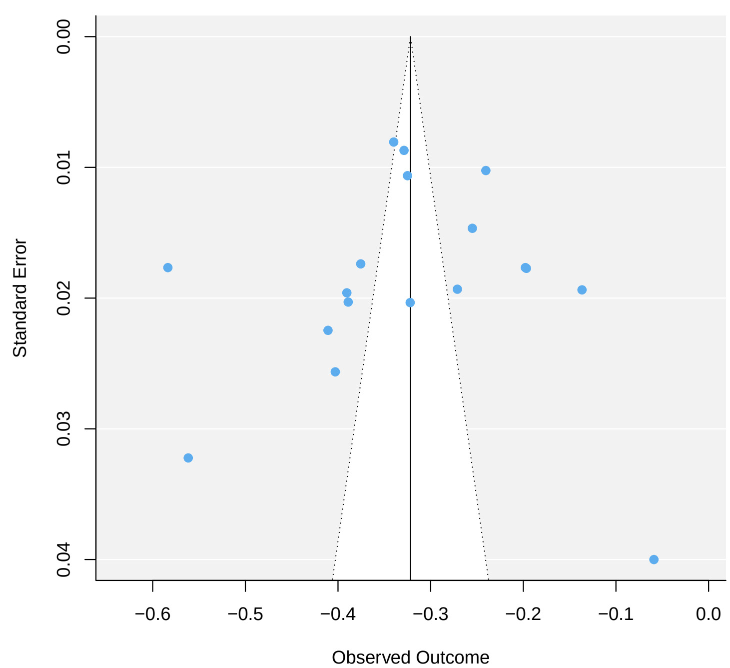

Figure A.7: Funnel plot: survey experiments

.

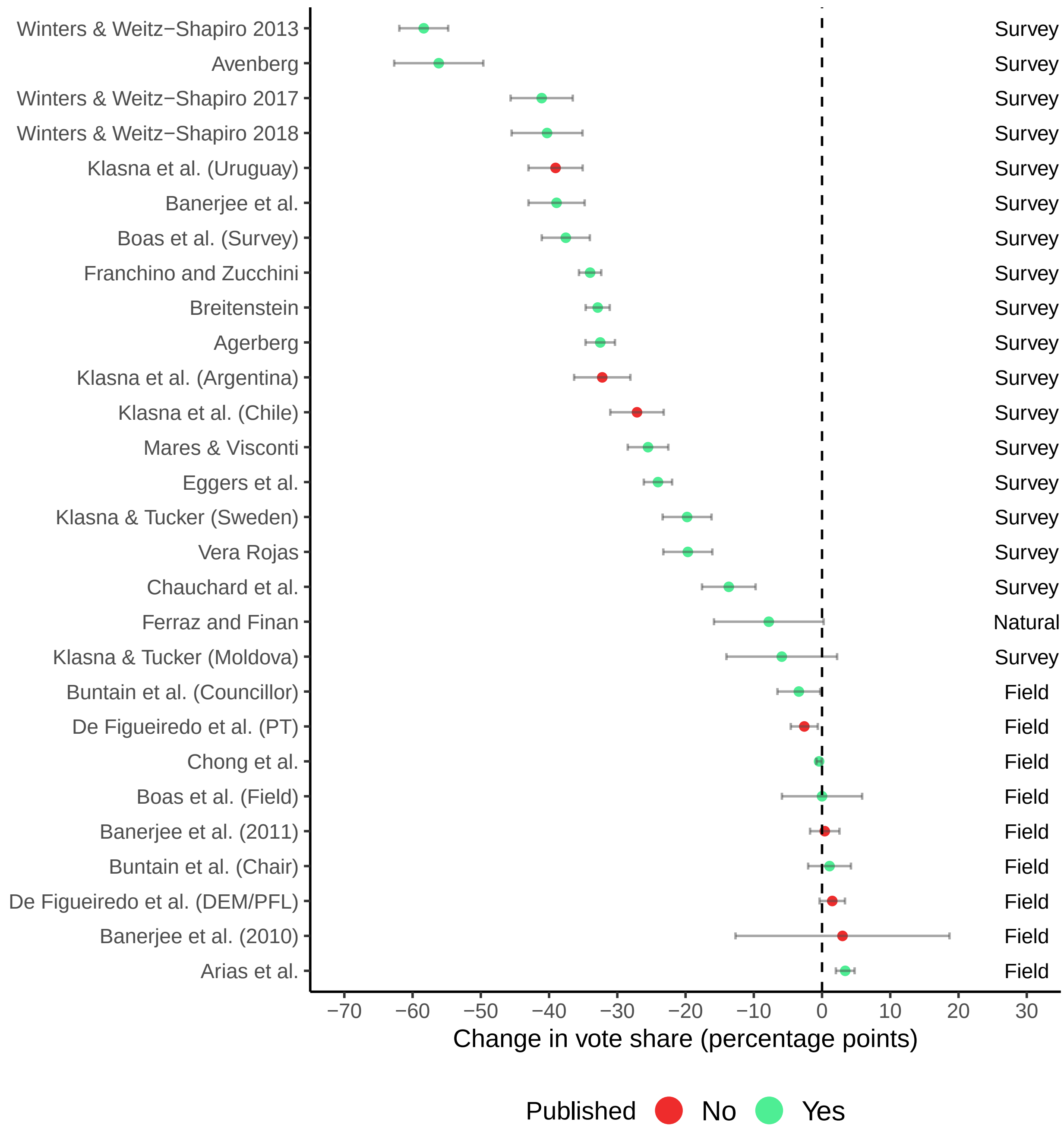

Figure A.8: All experiments by publication status: Average treatment effect of corruption information on vote share and 95% confidence intervals

.

A.6 Information quality

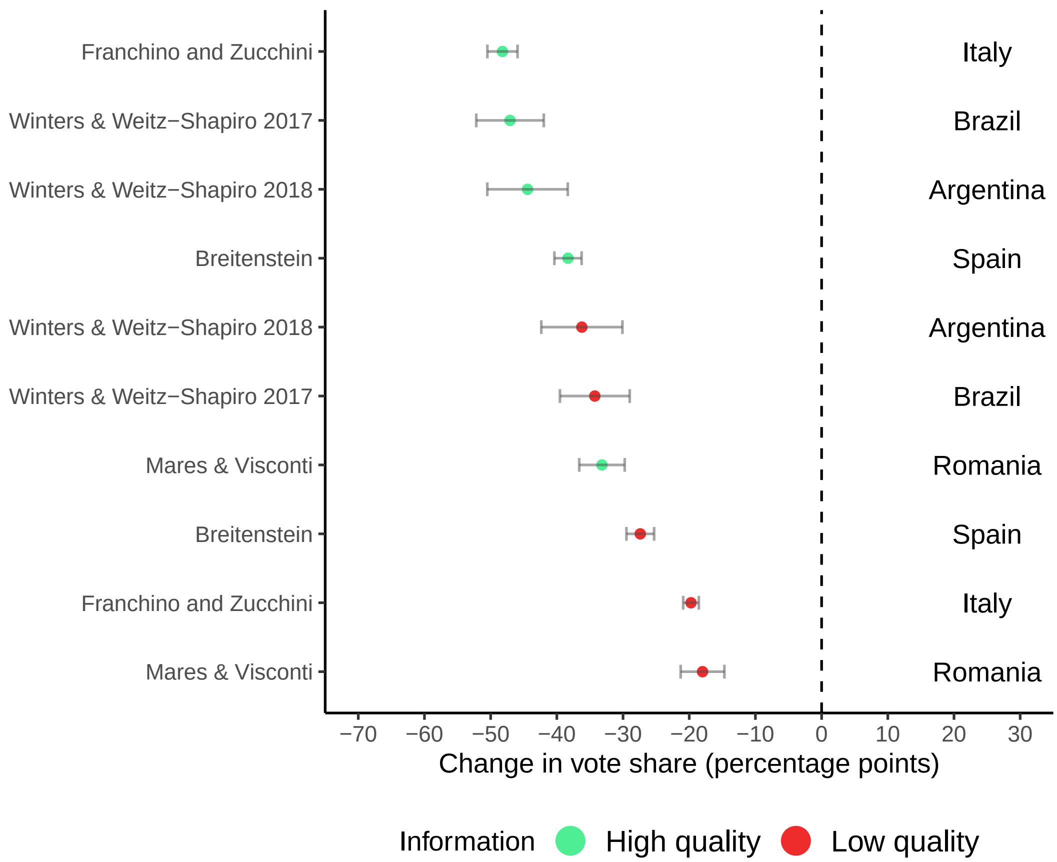

Figure A.9: Survey experiments by information quality: Average treatment effect of corruption information on vote share and 95% confidence intervals

.

A.7 Additional conjoint replications using predicted probabilities

A.7.1 Breitenstein (2019)

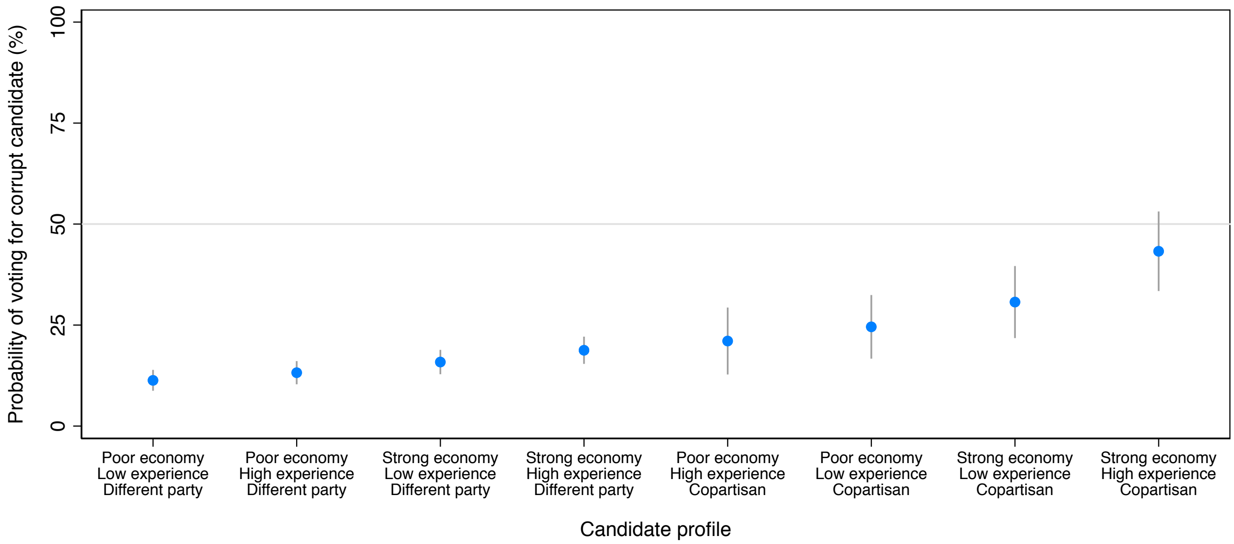

Figure A.10: Breitenstein (2019) conjoint: can the right candidate overcome corruption (clean challenger)?

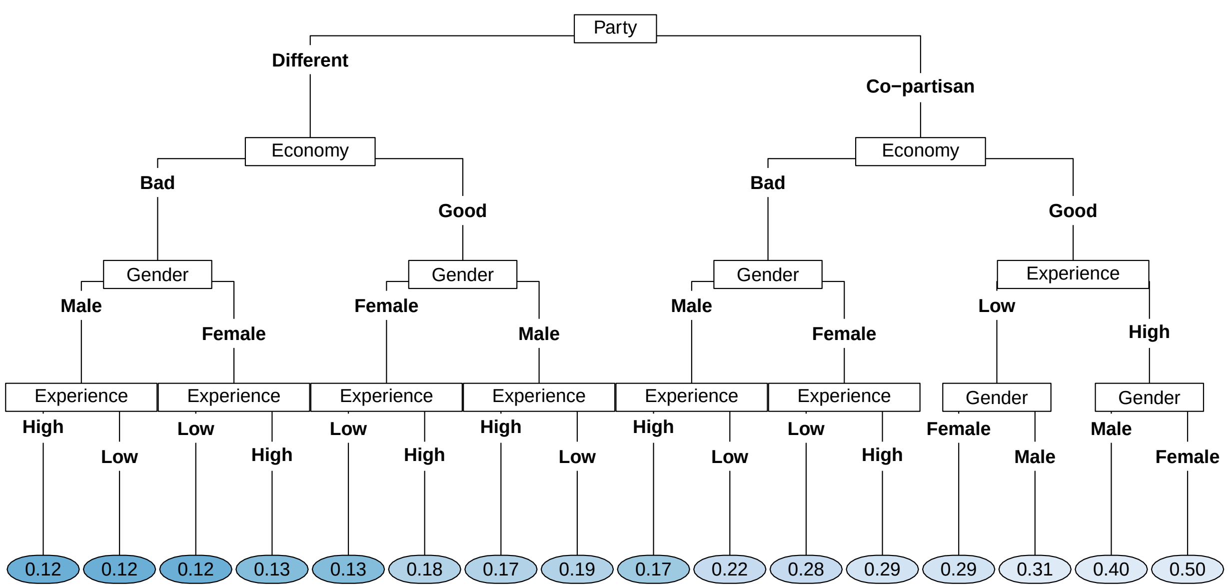

Figure A.11: Breitenstein (2019) conjoint decision tree: predicted probabilities of voting for corrupt politician with clean challenger

A.7.2 Franchino and Zucchini (2015)

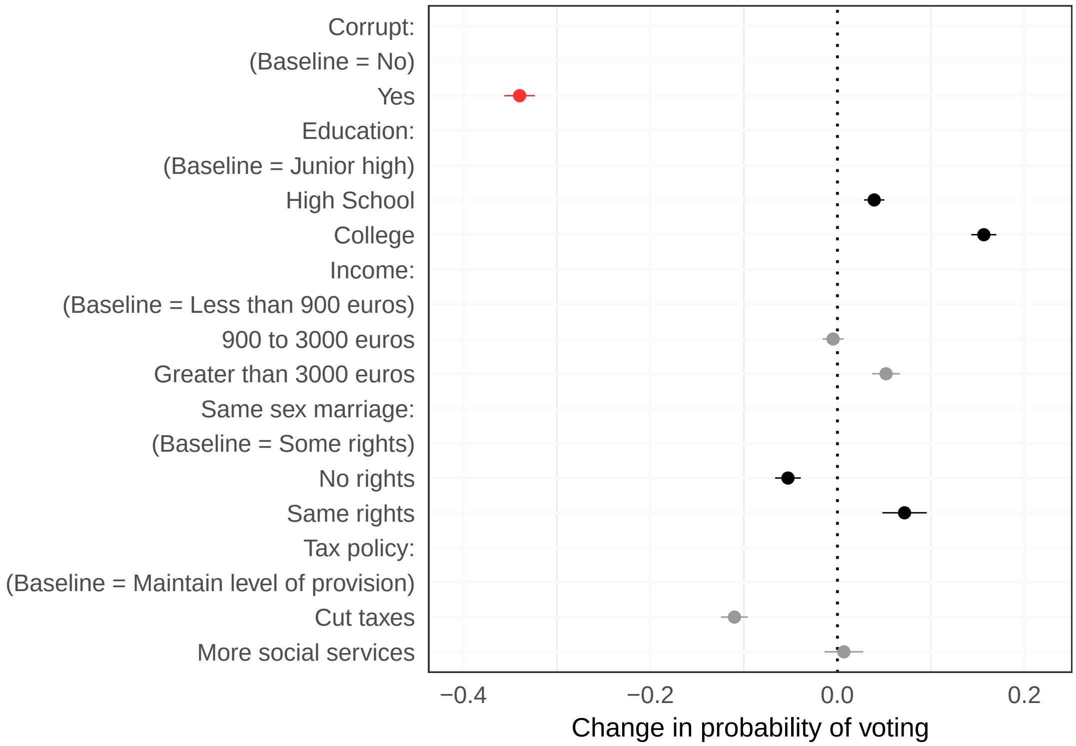

Figure A.12: Franchino and Zucchini (2015) conjoint: average marginal component effects

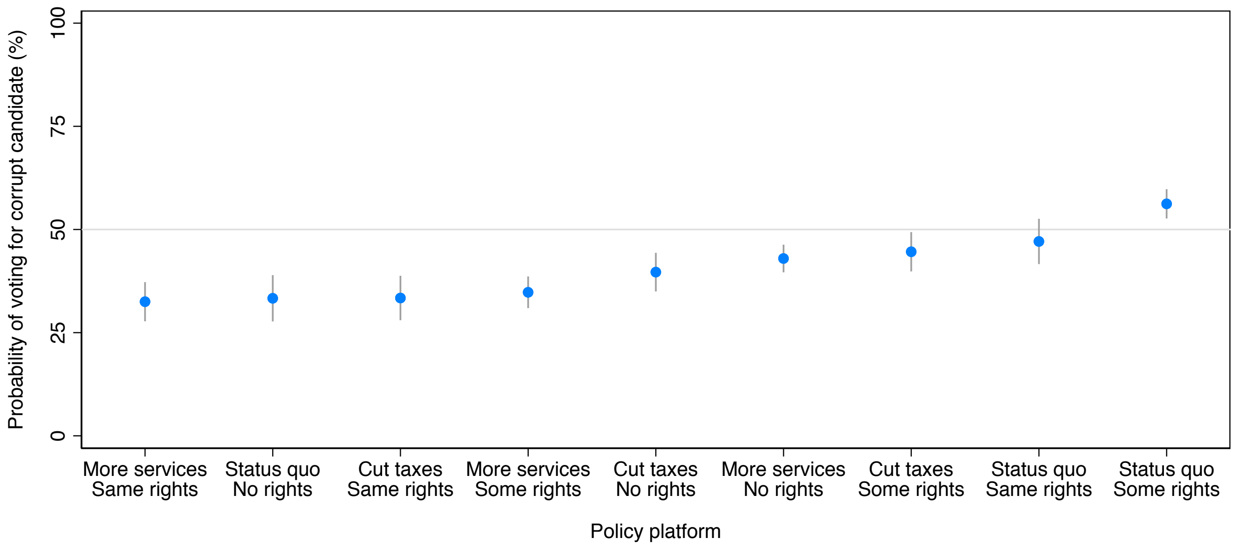

Figure A.13: Franchino and Zucchini (2015) conjoint: can policy positions overcome corruption (conservative respondents)?

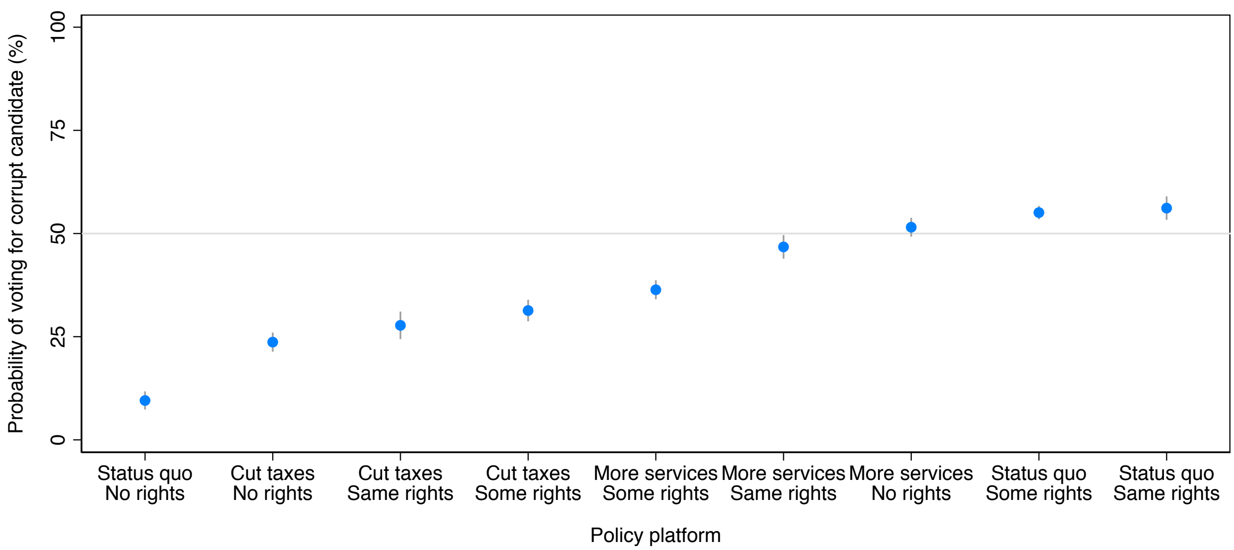

Figure A.14: Franchino and Zucchini (2015) conjoint: can policy positions overcome corruption (liberal respondents)?

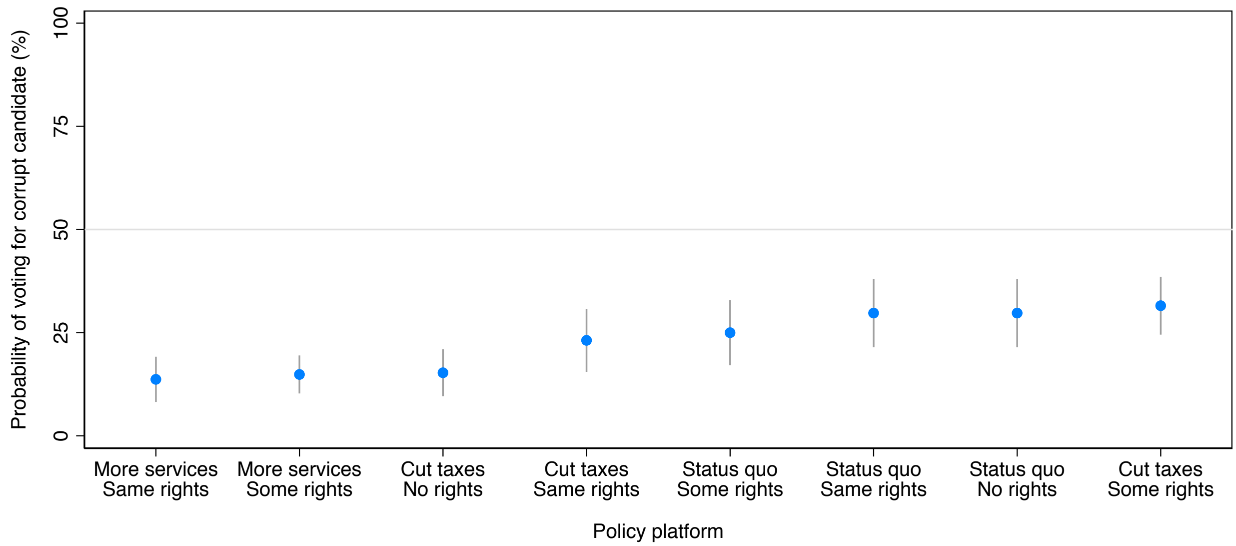

Figure A.15: Franchino and Zucchini (2015) conjoint: can policy positions overcome corruption (conservative respondents and clean challenger)?

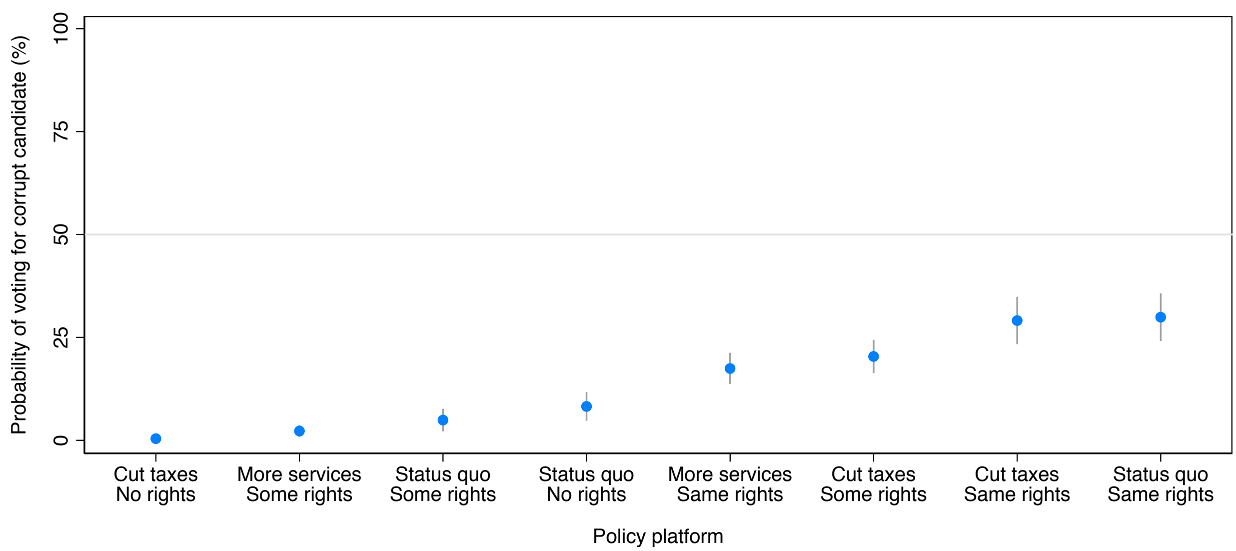

Figure A.16: Franchino and Zucchini (2015) conjoint: can policy positions overcome corruption (liberal respondents and clean challenger)?

A.8 Additional discussion of Boas, Hidalgo and Melo (2018)

Boas, Hidalgo and Melo (2018) find differential results in a pair of field and survey experiments conducted in Brazil—zero and null in field, large and negative in survey. They argue that this may reflect that norms against malfeasance in Brazil do not translate into action in real life. Boas, Hidalgo and Melo (2018) argue that "differences in research design are unlikely to account for much of the difference in effect size" and point to features specific to Brazil in their explanation of this discrepancy, namely lower salience of corruption to Brazilian voters in municipal elections and the strong effects of dynastic politics. However, meta-analysis demonstrates that this is not only the case for Boas, Hidalgo and Melo (2018)'s experiments in Brazil, but extends across a systematic review of all studies conducted to date. This suggests that the discrepancy between field and survey experimental findings is driven by differences in research design, rather than features specific to Brazilian voters.

First, if the differences reflected internalized norms, this would imply that even in hypothetical scenarios, if these issues were made salient in the experiment, they would have a large mitigating effect on the magnitude of the treatment effect. However, we see that the magnitude of treatment effects remains large in hypothetical scenarios even in the presence of mitigating factors. The difference in treatment effects therefore seems to reflect a mechanism akin to "hypothetical bias" found in stated preference surveys in environmental economics, in which respondents report a willingness to pay that is larger than what they will actually pay using their own money (Loomis 2011), rather than internalized norms.

Second, it should be noted that Boas, Hidalgo and Melo (2018) use an experimental design that mirrors that used by Weitz-Shapiro and Winters (2017) and Winters and Weitz-Shapiro (2018), rather than one that mirrors their own field experiment. That their estimates are not significantly different from those found in Weitz-Shapiro and Winters (2017) and Winters and Weitz-Shapiro (2018) is further evidence that the magnitude of these estimates may be a feature of that particular design.