Title: *Can Benefits Buy Forgiveness?* Social Welfare Programs and Voter Behavior Following a Corruption Shock
By: Trevor Norton

Question 1: *Describe a substantive question in social science. What theory are you (or the author of the paper you are replicating) assessing? Why should anyone care? (2 paragraphs)*

Vertical accountability is fundamental to the functioning of democracy (Schedler 1999). While democratic institutions are designed to enable voters to hold leaders accountable, this depends on citizens' willingness to act on available information. Democratic theory presumes that voters will sanction corrupt officials when presented with credible evidence. As Fearon (1999) explains, "[E]lections are seen as a sanctioning device that induces elected officials to do what the voters want. The anticipation of not being reelected in the future leads elected officials not to shirk their obligations to the voters in the present" (p. 56). Yet empirical work by scholars such as Boas, Hidalgo, and Melo (2018) shows that voters often fail to do so: even when confronted with clear evidence of misconduct, many continue to support incumbents. This challenges the assumption that information alone is sufficient to trigger electoral accountability. Why, then, do some voters punish corruption while others remain loyal?

This study examines whether material dependence on the state through social welfare benefits conditions political responses to credible corruption allegations. Specifically, it focuses on Brazil's Bolsa Família program, a large-scale conditional cash transfer (CCT) initiative aimed at reducing poverty, and Operation *Lava Jato* ("Car Wash"), a significant corruption investigation that implicated numerous political elites. The analysis examines whether Bolsa Família recipients, compared to non-recipients, are (a) less likely to withdraw support from political actors associated with corruption if they are viewed as guarantors of the program, and (b) more likely to perceive lower levels of corruption among those actors viewed as guarantors of the program. In doing so, the paper contributes to broader debates about democratic accountability and explores how economic dependence may shape and constrain political behavior.

This analysis engages with existing theories that suggest targeted social benefits can shape how individuals receive and respond to political information. In particular, it considers whether recipients of social welfare programs interpret credible allegations of corruption differently from non-recipients. Rather than immediately withdrawing support, beneficiaries may view these programs as part of an ongoing exchange with the state, potentially reinforcing loyalty to the politicians perceived as responsible for delivering them. While the theoretical mechanisms are explored in greater detail in the next section, the central question guiding this study is whether social welfare programs insulate the guarantors of these programs (that is, the political actors responsible for creating or maintaining them) from the electoral consequences of corruption.

Question 2: *The study you propose involves learning about a theory by observing certain of its implications. What one or two hypotheses that arise from the theory are you planning to assess (or does the author of the paper you are replicating assess)? Why or how does the theory justify your expectations about these hypotheses? (1 or 2 paragraphs)*

The theory developed in this study proposes that recipients of social welfare programs are politically insulated from the effects of corruption scandals and, as a result, are more likely to continue supporting the party responsible for implementing the program. This insulation arises from a combination of material dependence, symbolic attribution, and motivated reasoning. Social welfare benefits, particularly when they are visible and credibly attributed to specific political actors, are often interpreted as signals of care and responsiveness rather than neutral policy instruments (e.g., Hamel 2024; Manacorda, Miguel, and Vigorito 2011; Mettler 2002). As a result, beneficiaries may develop affective attachments to incumbent politicians, perceiving continued support as a means of protecting access to valued programs (see also Manacorda, Miguel, and Vigorito 2011). These attachments are further reinforced by directionally motivated reasoning, wherein individuals process new information in ways that preserve coherence between their material interests and prior political commitments (Kunda 1990; Taber and Lodge 2006). In this way, corruption allegations may be downplayed, rationalized, or reframed to maintain support for the perceived guarantors of the program.

From this theoretical foundation, two testable hypotheses emerge. First, recipients of social welfare programs perceive lower levels of corruption than non-recipients following corruption scandals (**H1**). This hypothesis reflects the idea that motivated reasoning may lead recipients to discount negative information that threatens the legitimacy of their political attachments. Second, recipients of social welfare programs reduce their support for incumbent parties less than non-recipients following corruption scandals (**H2**). This reflects the behavioral dimension of the theory: that material and symbolic ties to incumbents create durable political loyalty that persists even in the face of credible wrongdoing. Together, these hypotheses allow for an assessment of whether and how targeted social welfare programs shape the relationship between voters and elected officials in contexts marked by systemic corruption and weak institutional trust.[1]

Hypotheses

H1 (Corruption Perception Hypothesis): *Recipients of social welfare programs perceive lower levels of corruption than non-recipients following corruption scandals.*

H2 (Voting Behavior Hypothesis): *Recipients of social welfare programs reduce their support for incumbent parties less than non-recipients following corruption scandals.*

Question 3: *What data and research design will help you answer this question? Why are you making these choices? (Remember that a statistical model is not a research design.) (2 paragraphs)*

To answer whether social welfare programs insulate recipients from perceiving corruption (H1) and how this impacts their vote choice (H2), this study uses individual-level survey data from the 2014 and 2018 waves of the Comparative Study of Electoral Systems (CSES) in Brazil. These data are well-suited for the research question because they bracket a major corruption shock (the Lava Jato), allowing for comparison of political attitudes and behavior before and after its emergence. The surveys include direct measures of program receipt, vote choice, and corruption perceptions,[2] all of which are central to the proposed hypotheses. Moreover, the CSES's nationally representative sampling design ensures broad geographic and socioeconomic coverage, making the CSES particularly valuable for studying conditional political responses among beneficiaries of the Bolsa Família program.

To estimate the causal effect of program receipt, I employ a matched observational design using Mahalanobis distance optimal pair matching. Matching is appropriate in this context because it facilitates comparisons between recipients and non-recipients who are similar on observed characteristics, thereby reducing bias from confounding. I use Mahalanobis distance matching because it directly minimizes multivariate distance between units, preserving the original covariate structure without relying on a model to estimate treatment assignment. This approach follows Rosenbaum's (1989) argument that optimal pair matching improves balance by solving a global minimization problem. I also avoid propensity score matching (PSM), which, as King and Nielsen (2019) demonstrate, can increase

[1] For scholarship on systemic corruption and institutional trust, see Pavão (2018) and Adida et al. (2019).

[2] For the corruption perception variable, I am unable to compare pre- and post Lava Jato, as the necessary corruption perception variable does not exist in the 2014 dataset.

imbalance, inefficiency, and model dependence. Their critique shows that PSM approximates a completely randomized experiment, often degrading inference when data are already fairly balanced or when substantial pruning is involved.

What is more, I prefer matching over standard regression adjustment because it enables design-based control of confounding prior to outcome analysis, thereby reducing reliance on post hoc modeling assumptions and avoiding extrapolation across dissimilar units (Rosenbaum 2020). Ideally, a regression discontinuity design would have been more appropriate for identifying causal effects in this observational context. However, such an approach is not feasible due to data limitations—most notably, the absence of a measured variable indicating the number of children in the household, which precludes the construction of an eligibility threshold. Given this, I argue that Mahalanobis optimal pair matching offers the best preprocessing strategy for reducing covariate imbalance. The credibility of the design is further strengthened by leveraging an exogenous political shock (Lava Jato), which introduced plausibly independent variation in voters' exposure to credible allegations of corruption.

Question 4: *What are the advantages and disadvantages of this research design to addressing the substantive question? (2 paragraphs discussing both advantages and disadvantages of the research design; could be 1 paragraph for advantages and 1 for disadvantages or combined discussion across 2 paragraphs.)*

One major advantage of this research design is its ability to approximate experimental conditions within the constraints of observational data. By using matching, the design minimizes covariate imbalance between recipients and non-recipients of Bolsa Família, allowing for more credible estimation of causal effects than would be possible with simple regression models. The matching strategy is also strengthened by the use of a natural corruption shock, which provides a rare opportunity to evaluate voter responses to corruption in real-world conditions, outside of artificial experimental settings. Additionally, the design avoids reliance on model-dependent adjustments, instead prioritizing comparability between units through preprocessing (Rosenbaum 2020). Furthermore, the sensitivity analyses and simulation-based diagnostics further enhance credibility by testing the robustness of the results under plausible violations of identifying assumptions.

Nonetheless, the research design has several limitations, beginning with its observational nature. First, like all observational studies, it is vulnerable to unmeasured confounding. While matching improves covariate balance on observed variables, it cannot account for unobserved differences between treated and control groups that may also influence outcomes (though sensitivity analysis helps by quantifying how strong an omitted variable would need to be to overturn the conclusions). Second, the design relies on self-reported measures of program receipt and political attitudes, which may be subject to measurement error or social desirability bias—issues that are especially common in Latin American survey contexts (see Tourangeau and Yan 2007). Third, the cross-sectional structure of the data limits the ability to observe how attitudes change in response to political shocks, constraining inferences about the psychological mechanisms suggested by the theory. Fourth, because low income is one of the eligibility criteria for Bolsa Família, I include (self-reported) income as a matching variable. This poses challenges, though, as income itself may be influenced by program participation. Avoiding this endogeneity problem is difficult, as even potential proxies for low income (e.g., access to running water) are also likely affected by the program, which is designed to help households afford such basic necessities. Finally, while matching improves comparisons between beneficiaries and non-beneficiaries, a low-income non-beneficiary is not directly equivalent to a beneficiary. There are inherent differences that may explain why some low-income households do not receive benefits (for instance, lacking children or ineligibility due to criminal status).

Question 5: *Describe your measures and any indices you or the authors constructed. (1 paragraph)*

This study relies on several individual-level measures[3] from the 2014 and 2018 waves of the Brazilian CSES. The treatment variable, Bolsa Família receipt, is coded as a binary indicator (`bf_beneficiary`), where 1 indicates the respondent or someone in their household received benefits from the program in the past three years, and 0 indicates non-receipt; responses of "Don't know" or "No answer" are coded as missing. The first outcome (H1) variable, perceived corruption (`corruption_perceived`), is based on the question asking how widespread respondents believe corruption is in Brazil. Responses range from 1 ("Very widespread") to 4 ("Rarely happens"), with non-substantive responses treated as missing. Notably, this variable only appears in the 2018 data, so I am unable to employ a 2014 analysis for H1. The second outcome (H2) is vote choice, which was modified, becoming a binary indicator for support of the PT in the presidential runoff election (`vote_pt`), where 1 indicates a vote for the PT candidate (Dilma Rousseff in 2014, Fernando Haddad in 2018), and 0 includes all other responses. Like the others, null, blank, or nonresponse categories are coded as missing.

Matching covariates include sex (`male`), household income (`hh_income`), and education (`education`). Sex is coded as a binary indicator, where 1 = male and 0 =

[3] For more information about the coding of these variables, please see the codebook.

female. Household income is numeric, with responses of “Don’t know” or “No answer” treated as missing. Education level is taken from the education variable, which records completed schooling on a ten-point scale, ranging from no formal education (1) to a postgraduate degree (10). All numeric variables are standardized prior to matching to ensure comparability across scales.

Question 6: *Use data to make the case that your research design allows you to interpret observed quantities (like observed data comparisons or parameters of models fit to data) as theoretically relevant and clear: (Most people will only have to do either 6.1 and 6.2 or 6.3 and 6.4 here depending on whether you have a randomized design or an observational design). (relevance: i and ii)*
Part A: *If you are using an observational study design then explain how you will make the case for interpretable comparisons (this is the same as question as 'What is your identification strategy?'). That is, explain how you will use statistical adjustment (like matching or covariance adjustment aka "controlling for") to persuade yourself and others that the comparison that you are showing reflects what you say it does. If you are making comparative or causal inference, I assume you will explain the natural or quasi-experimental design and approach you will be using here. "I controlled for a lot of background variables in a linear model." will not be acceptable here. If you are making population inferences, you should explain your approach as well. (2 paragraphs plus some tables or figures)*

To estimate the causal effect of Bolsa Família receipt on perceived corruption and vote choice, I use a matched observational design that aims to approximate an as-if-randomized comparison. As already discussed, I implement Mahalanobis distance optimal pair matching using the `optmatch` package in `R`. This procedure pairs each treated unit (a Bolsa Família recipient) with a control unit (a non-recipient) that is most similar on a set of observed, pre-treatment covariates. For perceived corruption (H1), I match on education and household income, two characteristics that plausibly confound the relationship between treatment and attitudes toward corruption. Both socioeconomic status and educational attainment are strongly associated with patterns of political knowledge, media exposure, and tolerance for corruption in Latin America (Winters and Weitz-Shapiro 2013; Weitz-Shapiro and Winters 2017). For vote choice (H2), I additionally include sex in the matching process, as gender has been shown to influence patterns of partisanship and policy preferences in the Brazilian electorate (Agerberg 2014). The purpose of matching is to create a sample where treated and control units differ only on treatment status, allowing for comparisons that mimic the conditions of a randomized experiment. No regression adjustment is performed post-matching, as the identification strategy is rooted in pre-processing the data to eliminate covariate imbalance before estimating effects.

After matching, I estimate the average treatment effect on the treated (ATT) by calculating simple differences in group means. Covariate balance is assessed using the `RItools` package (Hansen and Bowers 2008), which provides randomization-based diagnostics. While the secondary survey data used here do not allow me to directly measure individual exposure to Lava Jato (e.g., whether respondents followed news about the investigation specifically), the credibility of the design is strengthened by the scope and salience of the scandal. Lava Jato was among the largest anti-corruption investigations in Brazilian history, generating widespread media coverage and implicating high-ranking politicians across multiple parties between 2014 and 2018 (e.g., Cioccari 2015). Because this shock was both nationally visible and external to the assignment of Bolsa Família benefits, it introduces plausibly exogenous variation in voters' exposure to credible political malfeasance. Taken together, the matched design, successful covariate adjustment, simulation-based inference, and incorporation of a high-salience, exogenous political event support a quasi-experimental identification strategy that enables credible and interpretable comparisons between treated and control groups.

Part B: *If you are using an observational study design, explain how you will judge the success of your adjustment strategy. For example, you may explain here about balance tests and other diagnostics that refer to the problem of adjustment for confounding or making the case for an as-if-randomized comparison, or an as-if-randomly sampled set of observations, etc.. (1 paragraph)*

To judge the success of my adjustment strategy and make the case for as-if-randomized comparisons, I employ a combination of covariate balance diagnostics, sensitivity analysis, and simulation-based performance checks. First, I evaluate covariate balance using standardized mean differences (SMDs) and omnibus tests from the `RItools` package (Hansen and Bowers 2008), ensuring that treated and control units are comparable on all matched covariates. Because balance is achieved through matching, I use a simple, unadjusted linear model to estimate the ATT,[4] avoiding additional covariate adjustments that would reintroduce model dependence. Second, I assess the risk of unobserved confounding using the `sensemakr` package (Cinelli and Hazlett 2020), which quantifies how strong an omitted variable would have to be, relative to observed covariates, to explain away the estimated treatment effect or render it statistically insignificant. Finally, I use simulation-based diagnostics to evaluate the overall performance of

[4] The R code for the unadjusted linear model is: `lm(outcome ~ treat + factor(pair_id))`.

the estimator. These include assessments of mean squared error (MSE), false positive rates, and statistical power, which indicate how precise and well-calibrated my inferences are under repeated sampling (these will be discussed later). Taken together, these diagnostics allow me to assess whether my adjustment strategy successfully reduces confounding and supports reliable, interpretable causal inference from observational data.

Balance statistics for hypothesis 1 are presented in Table 1, with the corresponding love plot shown in Figure 1. Balance statistics for hypothesis 2 are reported in Tables 2 and 3, with love plots in Figures 2 and 3. Across both hypotheses, the results show that the matching procedure was highly effective in reducing imbalance between treated and control groups. For hypothesis 1 (2018), both education and household income displayed substantial pre-matching differences (standardized differences of –0.376 and –0.475, respectively), which were reduced to near zero after matching (–0.009 and –0.052). For hypothesis 2, similar improvements are observed across both years: in 2014, male, education, and household income imbalances fell to 0.000, –0.012, and –0.007, respectively, while in 2018, the same covariates were reduced to 0.000, –0.025, and –0.074. In all cases, post-matching standardized differences fell well below conventional thresholds, indicating that covariates were successfully balanced and that the matched samples provide a credible basis for estimating treatment effects.

Table 1: Balance Table - Hypothesis 1 (2018)

Variable	Pre-Matching Mean (Control)	Pre-Matching Mean (Treated)	Pre-Matching Std. Diff	Post-Matching Mean (Control)	Post-Matching Mean (Treated)	Post-Matching Std. Diff
Education	5.151	4.309	–0.376	4.329	4.309	–0.009
HH Income	3,012.991	1,618.520	–0.475	1,770.827	1,618.520	–0.052

Figure 1: Love Plot - Hypothesis 1 (2018)

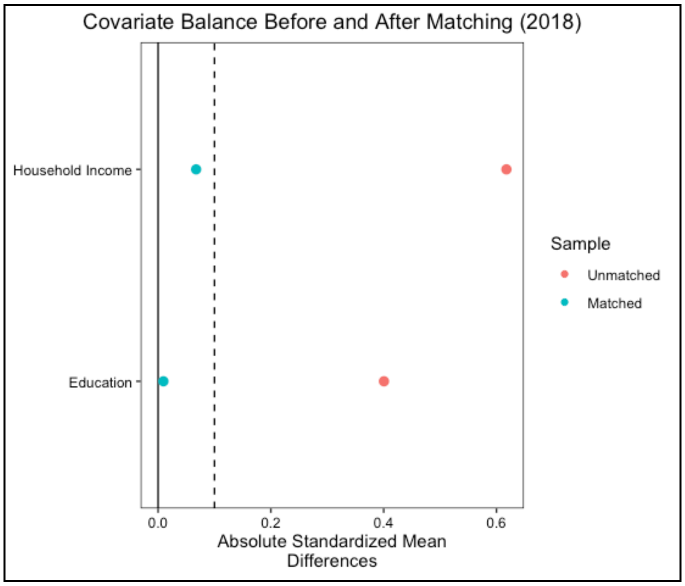


Table 2: Balance Table - Hypothesis 2 (2014)

Variable	Pre-Matching Mean (Control)	Pre-Matching Mean (Treated)	Pre-Matching Std. Diff	Post-Matching Mean (Control)	Post-Matching Mean (Treated)	Post-Matching Std. Diff
Male	0.516	0.464	–0.104	0.464	0.464	0.000
Education	4.673	3.851	–0.362	3.878	3.851	–0.012
HH Income	2731.793	2003.124	–0.230	2025.788	2003.124	–0.007

Table 3: Balance Table - Hypothesis 2 (2018)

Variable	Pre-Matching Mean (Control)	Pre-Matching Mean (Treated)	Pre-Matching Std. Diff	Post-Matching Mean (Control)	Post-Matching Mean (Treated)	Post-Matching Std. Diff
Male	0.535	0.385	–0.302	0.385	0.385	0.000
Education	5.117	4.371	–0.328	4.428	4.371	–0.025
HH Income	3,071.521	1,579.000	–0.484	1,806.475	1,579.000	–0.074

Figure 2: Love Plot - Hypothesis 2 (2014)

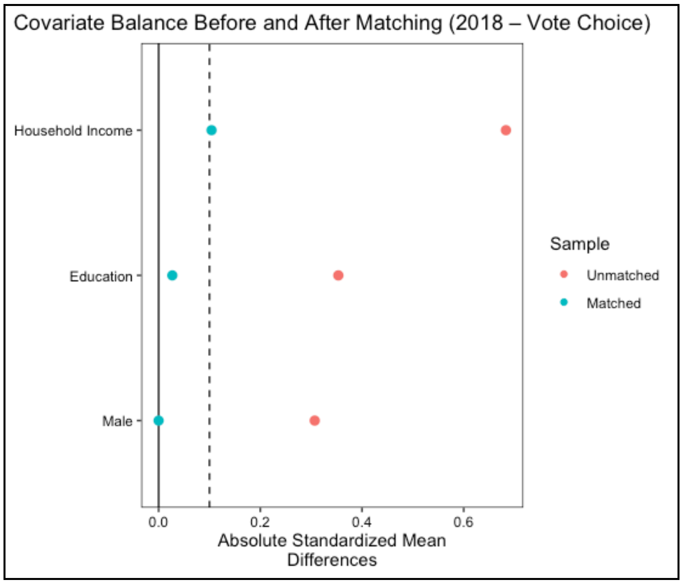
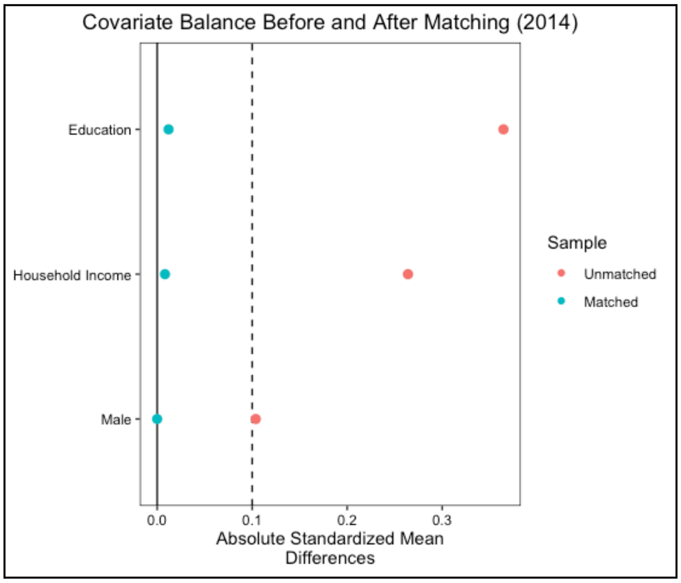


Figure 3: Love Plot - Hypothesis 2 (2018)



Question 7: *Explain your plans for any missing data or extreme outcome or covariate values you may encounter when you get the real data (or perhaps you have the background data but not the real outcomes, so you can explain your plans for such data issues in that case here too). (1 or 2 paragraphs)*

When working with the CSES data, I anticipate some missing values in covariates such as income, education, and sex, as well as potential nonresponse on outcome variables like vote choice or perceived corruption. For covariates used in matching, I use listwise deletion (by default, since the matching procedure requires complete cases to compute pairwise distances). As a result, all units with missing values on matched covariates are automatically excluded from the analysis before estimation, and the linear models I use are fit only to the already-complete matched sample. For outcome variables, I also apply listwise deletion by dropping any matched observations with missing outcome values. Before doing so, I will assess whether missingness is plausibly random by creating a missingness indicator (e.g., `is.na(vote_pt)`) and checking whether it is significantly associated with treatment status or covariates using chi-square tests and logistic regression. If I find evidence of systematic missingness, I will report the observed differences between respondents and nonrespondents and note the potential for selection bias.

For extreme values in this analysis, I do not exclude them outright, as such observations may still represent substantively valid cases. Instead, I flag values that fall more than three standard deviations from the mean and document their distribution across treatment groups. This approach allows me to acknowledge and track potential outliers without removing them from the analysis or altering the matching procedure. In the actual results, the flagged cases are relatively rare: in 2018, only 24 non-beneficiaries and two beneficiaries fell into the extreme range, while in 2014, 24 non-beneficiaries and 8 beneficiaries were flagged. This indicates that the presence of extreme household income values is limited across treatment groups, reducing concern that they disproportionately influence the estimated treatment effects.

Question 8: *What statistical tests do you plan to use? Explain why you chose these tests and any decision making criteria you will use upon seeing the results of the tests. You should also engage with the problem of multiple testing here if you are going to show the results of more than one test. (Recall that confidence intervals and hypothesis tests convey more or less the same information. So a confidence interval is a form of testing.) (1 paragraph)*

My primary statistical tests estimate the ATT using matched linear models with fixed effects for matched pairs. I conduct randomization inference by permuting treatment assignments within matched pairs to generate permutation-based p-values under the sharp null hypothesis of no treatment effect. I also report confidence intervals and p-values from linear models for both vote choice and perceived corruption, interpreting them alongside the permutation results. To assess estimator performance and sampling variability, I conduct power simulations and calculate MSE under both design-based and noise-added conditions. I also use the `sensemakr` package to evaluate the robustness of estimates to potential unobserved confounding. To compare results across years for

H2, I calculate a *z*-score for the difference in ATT estimates between 2014 and 2018, which provides a direct test of whether the observed effects differ significantly across survey waves. Because I test two main outcomes and conduct additional robustness checks, I address the risk of multiple testing by focusing on the consistency of effect direction, magnitude, and sensitivity rather than applying formal family-wise error corrections.

Question 9: *Explain how you will judge the performance of those tests. Will you only use the simple false positive rate and power? Or do you need to add family-wise error rate? False discovery rate? Or something else? Explain why you made this choice. (1 paragraph)*

To judge the performance of my statistical tests, as noted in the previous question, I use simulation-based diagnostics, randomization inference, and sensitivity analysis. I assess false positive rates, statistical power, and MSE by simulating treatment effects of varying magnitudes and re-estimating effects across resampled matched datasets. I also generate a power curve in base R to visualize how likely my design is to detect treatment effects of different sizes. For hypothesis testing, I use the `RItools` package to conduct randomization inference, calculating permutation-based p-values under the sharp null within the matched design. I use the `sensemakr` package to assess sensitivity to unobserved confounding, estimating how strong an omitted variable would need to be to eliminate the observed effect. Since I test only two pre-specified outcomes, I do not apply family-wise error rate corrections.

Question 10: *Show and explain how your test performs in regards those properties (at least you will show false positive rate and power). (2--4 paragraphs)*

For H1 (perceived corruption), the estimated false positive rate was 6.1 percent, slightly above the nominal 5 percent threshold. Under noise-free conditions, the estimator exhibited low MSE (.002) and negligible bias (< .001). When random noise was introduced to simulate realistic variation in outcomes, the MSE decreased to .001 and bias remained near zero (.002). These results suggest that the estimator is mostly stable, well-calibrated, and unlikely to generate spurious significance due to sampling variability or imbalance.

However, the estimator performs poorly in terms of power for H1. Power curves generated by simulating a range of true treatment effects (τ = 0 to 0.50) reveal that the model has very limited capacity to detect small-to-moderate effects in perceived corruption evaluations. Under realistic outcome noise, power was just 12.3 percent for a true effect of 0.05 and only reached 79.6 percent at τ = 0.10. The 80 percent threshold (which is commonly used as a benchmark for adequate power) was not surpassed until τ = 0.15, and full power (≥ 99 percent) was only achieved for effect sizes of 0.20 and above. These results suggest that while the H1 estimator is unbiased, it may be underpowered for detecting substantively smaller effects. A power curve graph can be found in Figure 4, and a table in Table 4.

Figure 4: Power Curves - H1 (2018)

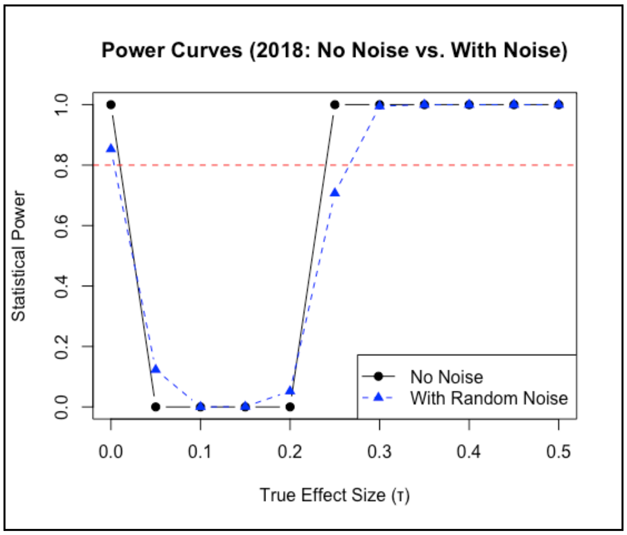


Table 4: Power Curves - H1 (2018)

0.00	1	.853
0.05	0	0.123
0.10	0	0.000
0.15	0	0.001
0.20	0	0.052
0.25	1	0.707
0.30	1	0.995
0.35	1	1.000
0.40	1	1.000
0.45	1	1.000
0.50	1	1.000

For H2 (vote choice), the estimator also performs well in terms of false positive rate and precision. In 2014, the false positive rate was 3.3 percent, and in 2018 it was 4.6 percent, both below the nominal 5 percent threshold. The estimator yielded low MSE and minimal bias across both years. In 2014, the MSE was .001 with bias = –.001 under ideal conditions; with added noise, MSE declined to < .001 and bias dropped to < .001. In 2018, MSE was .001 (bias = .002) in the no-noise model and .001 (bias = –.001) under realistic conditions. These findings confirm that the ATT estimates are statistically precise and robust across both waves.

Power curves for H2 outcomes demonstrate much stronger performance than for H1. In 2014, the estimator had nearly full power to detect effects of 0.10 (97.5 percent) and achieved 100 percent power for τ ≥ 0.15. The 2018 model had more modest power: only 10 percent at τ = 0.05 and 52 percent at τ = 0.15, but it reached 70.7 percent at τ = 0.25 and 99.5 percent at τ = 0.30. These results indicate that the estimator is highly sensitive to moderate-to-large effects in vote choice outcomes, particularly in 2014, while the 2018 vote model is slightly less powerful but still capable of detecting substantial shifts in support. Overall, the estimator's power performance reinforces confidence in its ability to recover meaningful treatment effects when they exist—especially in behavioral, rather than perceptual, outcomes. Power curve graphs and tables can be found in Figures and Tables 5 and 6.

Figure 5: Power Curves - H2 (2014)

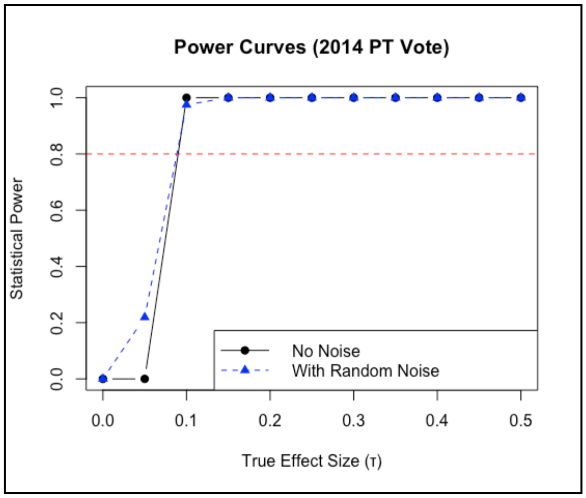


Figure 6: Power Curves - H1 (2018)

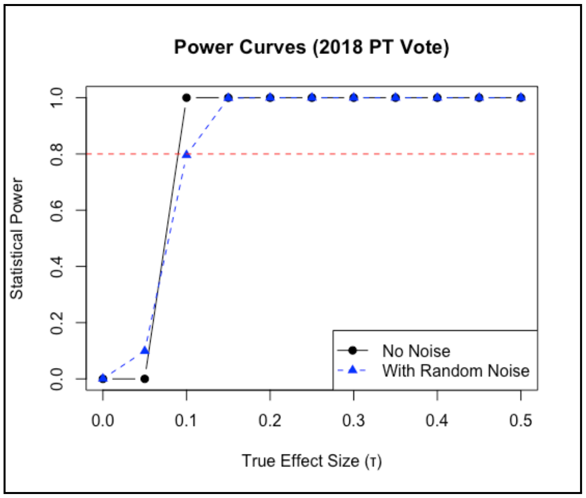


Table 5: Power Curves - H2 (2014)

0.00	0	0.000
0.05	0	0.219
0.10	1	0.975
0.15	1	1.000
0.20	1	1.000
0.25	1	1.000
0.30	1	1.000
0.35	1	1.000
0.40	1	1.000
0.45	1	1.000
0.50	1	1.000

Table 6: Power Curves - H2 (2018)

0.00	0	0.001
0.05	0	0.099
0.10	1	0.796
0.15	1	0.999
0.20	1	1.000
0.25	1	1.000
0.30	1	1.000
0.35	1	1.000
0.40	1	1.000
0.45	1	1.000
0.50	1	1.000

With regards to sensitivity, the results point to a clear difference across the two hypotheses. For H1 (Figure 7), the robustness values are very small: unobserved confounders explaining as little as 1 to 2 percent of the residual variance in both treatment and outcome would be sufficient to reduce the effect estimate to zero or render it statistically insignificant. This indicates that the corruption perception result is highly fragile to potential hidden bias. In contrast, the ATT estimates for H2 are much more robust. For the 2014 analysis (Figure 8), the sensitivity to unobserved confounders is moderate: eliminating the effect would require confounders explaining over 13 percent of the residual variance, while undermining statistical significance would take just over 5 percent. The 2018 result (Figure 9) is stronger, with robustness values suggesting that confounders would need to explain nearly 24 percent of the residual variance to reduce the effect to zero, and over 16 percent to make it statistically insignificant. Taken together, these diagnostics suggest that while the corruption perception effect is not robust to unobserved confounding, the electoral effects of Bolsa Família (especially in 2018) are far more stable and unlikely to be explained away by omitted variables.

Figure 7: Sensitivity Analysis - H1 (2018)

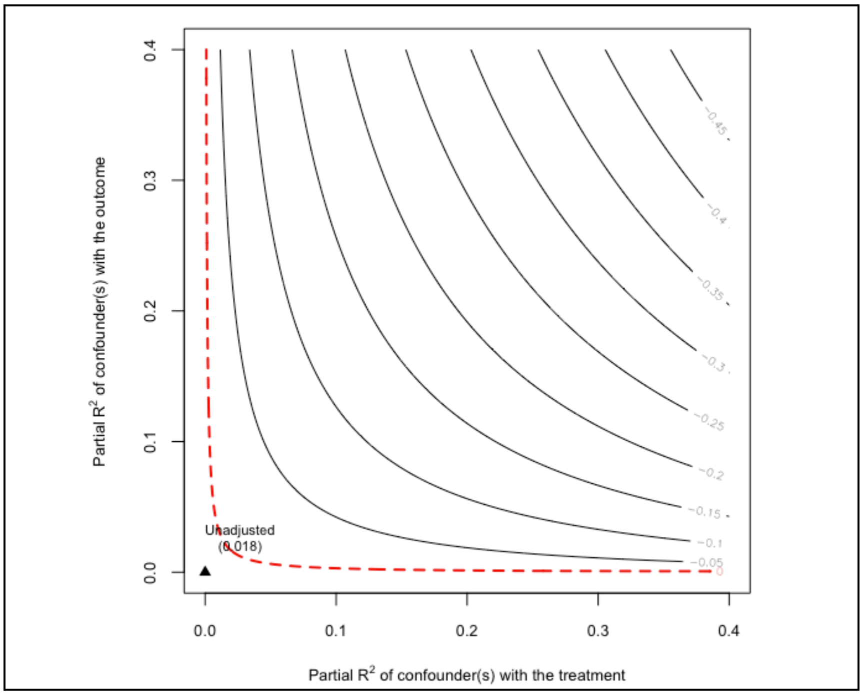


Figure 8: Sensitivity Analysis - H2 (2014)

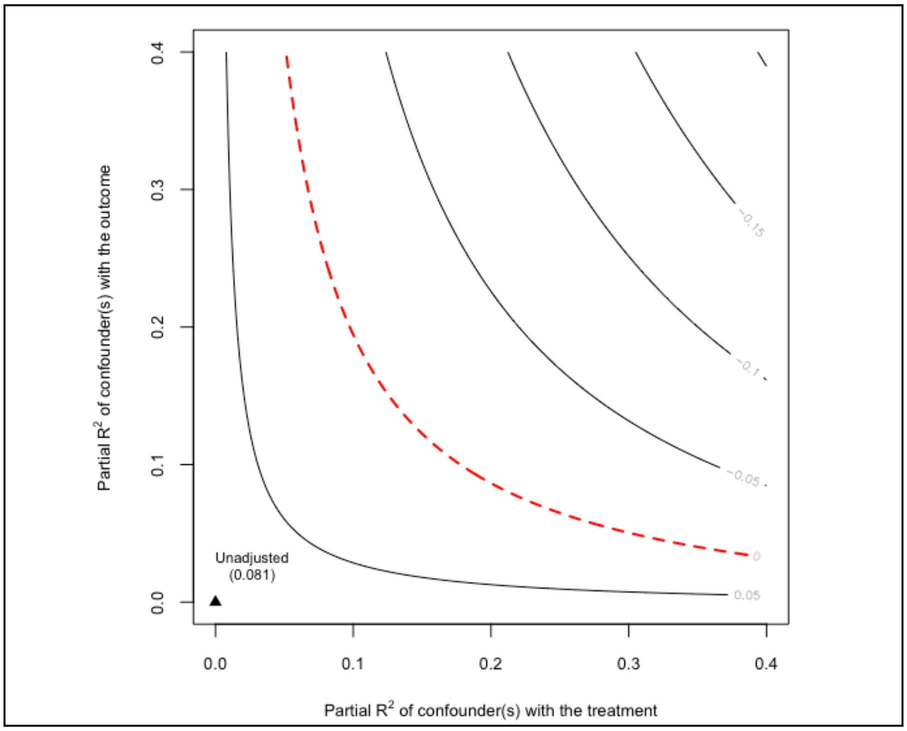
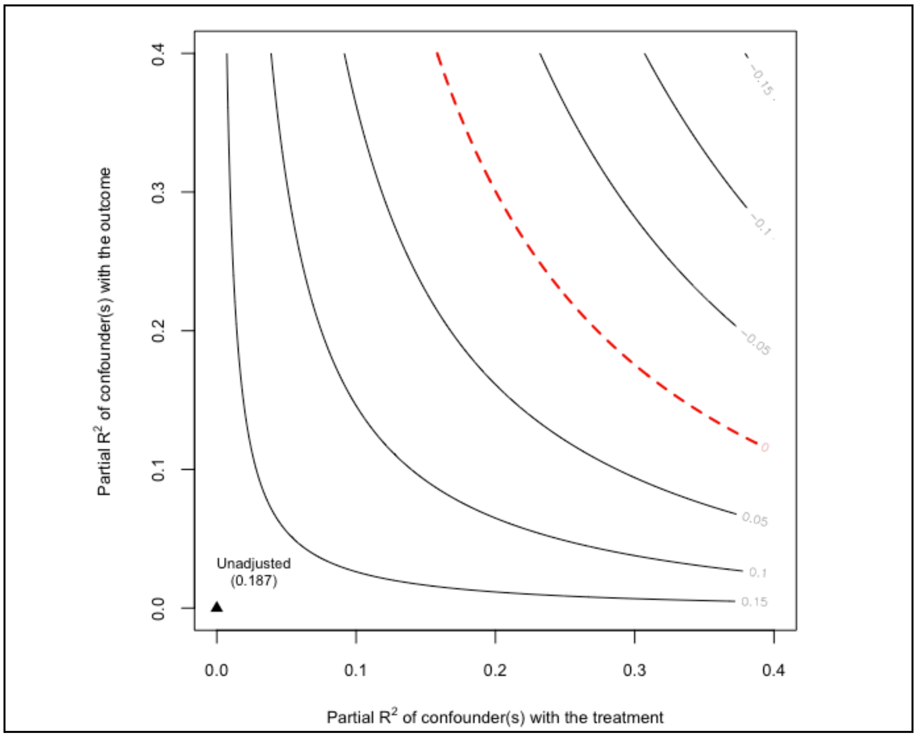


Figure 9: Sensitivity Analysis - H2 (2018)



Question 11: What statistical estimators do you plan to use? Explain why you chose these estimators. Especially explain what is your target of estimation --- what is the estimand? (1 paragraph)

To estimate the causal effect of Bolsa Família receipt on political attitudes and behavior, I use the ATT as my main estimand. The ATT represents the difference in expected outcomes (either perceived corruption (H1) or incumbent vote choice (H2)) between Bolsa Família recipients and the counterfactual outcomes those same individuals would have experienced had they not received the benefit. This estimand is appropriate given the study's focus on how actual recipients responded to the Lava Jato scandal. For the voter behavior outcome (H2), I also treat the difference in ATT estimates between 2014 and 2018 as an estimand of interest, assessed using *z*-scores to determine whether the change across years is statistically significant. I estimate the ATT using simple differences in means, based on a linear model applied to matched data constructed via Mahalanobis distance optimal pair matching. I include fixed effects for matched pairs and do not apply additional covariate adjustments post-matching.

Question 12: *Explain how you will judge the performance of those estimators (especially bias and MSE)?* (1 paragraph)

To judge the performance of my estimator, I assess bias and MSE for both main outcomes: perceived corruption (H1) and vote choice (H2). For each outcome and year, I begin by removing the estimated treatment effect from the observed outcome to construct a baseline version that reflects no treatment effect. Then, I simulate new outcomes by adding back in the known ATT to treated units, either deterministically (to assess design-based variation) or with added random noise based on the model's residual variance (to reflect sampling uncertainty). Across 1,000 simulations, I re-estimate the ATT using the same matched-pair linear model applied in the main analysis. I calculate bias as the difference between the average of the simulated estimates and the true ATT that was added during simulation. MSE is calculated as the average squared difference between each simulated estimate and the true ATT.

Question 13: *Show and explain how your estimator performs in regards those properties (at least bias and MSE). (2--4 paragraphs)*

To assess how well the estimator recovers the true treatment effect, I conducted simulation-based diagnostics measuring bias and MSE for all three models: perceived corruption in 2018 (H1), and PT vote choice in both 2014 and 2018 (H2). Bias represents the average deviation between the estimated treatment effect and the true value used in the simulation, while MSE captures the average squared error and incorporates both bias and variance. An estimator with low bias and MSE is considered reliable and well-calibrated.

For H1 (corruption perceptions in 2018), the estimator performs well under both ideal and realistic conditions. In the no-noise scenario, the bias was negligible (<.001), and MSE was .002. When random noise was added to reflect real-world outcome variation, bias remained low (.002), and MSE decreased to .001. These results indicate that the estimator does not systematically over- or underestimate the treatment effect and is robust to sampling variability.

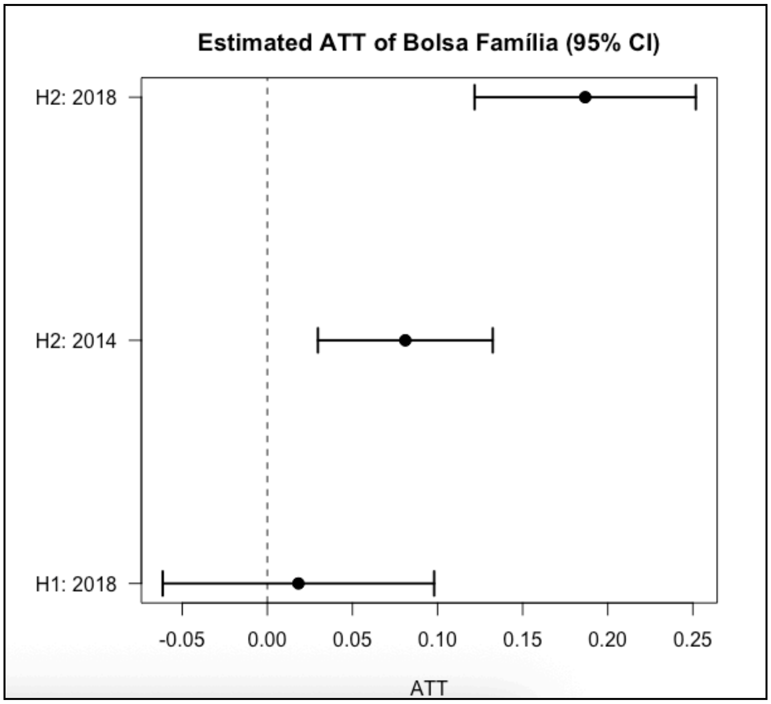
For H2, the estimator was similarly well-behaved in both waves. In 2014, bias under design-based uncertainty was –.001 with an MSE of .001. Under realistic noise conditions, bias dropped to virtually zero (<.001) and MSE to <.001. In 2018, bias under no-noise conditions was slightly higher (.002), with an MSE of .001. Yet when noise was introduced, bias reversed sign but remained small (–.001), and MSE declined to .001. Across all specifications, bias stayed close to zero, and MSE remained low, suggesting that the estimator performs well even in finite samples.

Taken together, these simulation results provide evidence that the matched estimator is both unbiased and precise. The bias values are close to zero in all conditions, and the MSE remains low across outcomes and waves. This reinforces confidence in the reliability of the estimated treatment effects reported in the main analysis, and supports the conclusion that the estimator is not vulnerable to major distortion from design-based or random variation.

Question 14: *Make one mock figure or table of the kind you plan to make when you use the actual outcome. Interpret the results of the mock analysis as if it were the real analysis. Saying something like, "If the real outcome were as I have simulated it, then the following table/figure would mean such and so about the theory." (1 paragraph)*

Given that I already have the data (and results), in Figure 10, I will provide the estimated ATTs (with 95 percent confidence intervals) for the two hypotheses. For H1 (perceived corruption in 2018), the effect of Bolsa Família is substantively small (ATT ≈ 0.02) and the confidence interval crosses zero, indicating no detectable difference in corruption perceptions between recipients and non-recipients. For H2 (vote choice), however, the results point to a consistent and sizable electoral effect: in 2014, the ATT is about 0.08 (≈8 percentage points more likely to vote for the PT), while in 2018 the ATT rises to about 0.19 (≈19 points). Both estimates are statistically distinguishable from zero, indicating that the treatment effect is reliably different from no effect in both years. Importantly, a *z*-test comparing the two coefficients shows that the 2018 effect is significantly larger than the 2014 effect (ATT = 0.106, SE = 0.042, z = 2.50, p = 0.012). This result suggests that the treatment's impact not only persists but grows over time, with the 2018 estimate reflecting a substantively stronger association than the one observed in 2014.

Figure 10: Estimated ATTs for H1 and H2



Taken together, these results suggest that while Bolsa Família did not meaningfully alter how recipients perceived corruption, it substantially shaped their voting behavior, with its electoral impact strengthening rather than weakening in the wake of Lava Jato. This finding is noteworthy because it implies that even though recipients may have been aware of corruption, their material dependence on the program outweighed these concerns when casting their votes, reinforcing electoral support for the PT rather than diminishing it. Of course, this has implications for democratic accountability, as it raises the possibility that material benefits can insulate incumbents from electoral punishment even in the face of major corruption scandals. In this sense, social policy may not only redistribute resources but also reshape the incentives voters face when weighing performance and integrity at the ballot box.

Question 15: Include a code appendix and a link to the github repository for this paper.
[*completed*]

Question 16: References (APSA FORMAT)

Agerberg, Mattias. 2014. "Perspectives on Gender and Corruption." *QOG THE QUALITY OF GOVERNMENT INSTITUTE*. https://gupea.ub.gu.se/handle/2077/38887.

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* 63(2): 385–400. doi:10.1111/ajps.12413.

Cinelli, Carlos, and Chad Hazlett. 2020. "Making Sense of Sensitivity: Extending Omitted Variable Bias." *Journal of the Royal Statistical Society Series a (Statistics in Society)*. https://pubag.nal.usda.gov/catalog/6819496.

Cioccari, Deysi. 2015. "OPERAÇÃO LAVA JATO: ESCÂNDALO, AGENDAMENTO e ENQUADRAMENTO." *Revista Alterjor* 12(2): 58–78. http://www.unigran.br/mercado/paginas/arquivos/edicoes/9/6.pdf.

Fearon, James D. 1999. "Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance." In *Cambridge University Press eBooks*, , 55–97. doi:10.1017/cbo9781139175104.003.

Hamel, Brian T. 2024. "Traceability and Mass Policy Feedback Effects." *American Political Science Review*: 1–16. doi:10.1017/s0003055424000704.

Hansen, Ben B., and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23(2). doi:10.1214/08-sts254.

King, Gary, and Richard Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27(4): 435–54. doi:10.1017/pan.2019.11.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108(3): 480–98. doi:10.1037/0033-2909.108.3.480.

Manacorda, Marco, Edward Miguel, and Andrea Vigorito. 2011. "Government Transfers and Political Support." *American Economic Journal Applied Economics* 3(3): 1–28. doi:10.1257/app.3.3.1.

Mettler, Suzanne. 2002. "Bringing the State Back in to Civic Engagement: Policy Feedback Effects of the G.I. Bill for World War II Veterans." *American Political Science Review* 96(02): 351–65. doi:10.1017/s0003055402000217.

Rosenbaum, Paul. 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84(408): 1024–32. doi:10.1080/01621459.1989.10478868.

Rosenbaum, Paul. 2020. "Modern Algorithms for Matching in Observational Studies." *Annual Review of Statistics and Its Application* 7(1): 143–76. doi:10.1146/annurev-statistics-031219-041058.

Schedler, Andreas. 1999. “Conceptualizing Accountability.” In *Lynne Rienner Publishers eBooks*, , 11–28. doi:10.1515/9781685854133-003.

Taber, Charles S., and Milton Lodge. 2006. “Motivated Skepticism in the Evaluation of Political Beliefs.” *American Journal of Political Science* 50(3): 755–69. doi:10.1111/j.1540-5907.2006.00214.x.

Tourangeau, Roger, and Ting Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin* 133(5): 859–83. doi:10.1037/0033-2909.133.5.859.

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. doi:10.1086/687287.

Winters, Matthew S., and Rebecca Weitz-Shapiro. 2013. “Lacking Information or Condoning Corruption: When Do Voters Support Corrupt Politicians?” *Comparative Politics* 45(4): 418–36. doi:10.5129/001041513806933583.