Supporting Information Appendix for

Political Power and Reciprocity

Daniel Enemark	Clark C. Gibson	Mathew D. McCubbins	Brigitte Seim
Enemark Consulting	UC San Diego	Duke University	UNC Chapel Hill

Table of Contents

Reciprocity is central to our understanding of politics. Most political exchanges—whether they involve legislative vote-trading, inter-branch bargaining, constituent service, or even the corrupt exchange of public resources for private wealth—require reciprocity. But how does reciprocity arise? Do government officials learn reciprocity while holding office, or do recruitment and selection practices favor those who already adhere to a norm of reciprocity? We recruit Zambian politicians who narrowly won or lost a previous election to play behavioral games that provide a measure of reciprocity. This combination of regression-discontinuity and experimental designs allows us to estimate the effect of holding office on behavior. We find that holding office increases adherence to the norm of reciprocity. This is the first study to identify causal effects of holding office on politicians' behavior. In this Appendix, we discuss supplementary information about the study.

www.pnas.org/cgi/doi/10.1073/pnas.1511501113

S1. Regression Discontinuity Design (RDD)

RDDs exploit the fact that some treatments are applied according to a rule, where cases are assigned to treatment when they cross a cutpoint in a continuous "forcing variable." In a successful RDD, the treated and untreated groups are essentially identical (pre-treatment) within the immediate vicinity of the cutpoint. Under these circumstances, potential outcomes are (conditioned on the forcing variable) independent of assignment to treatment, allowing us to estimate average treatment effects within the vicinity of the cutpoint. This study uses an electoral RDD, with margin of victory (the forcing variable) determining who obtains office (the treatment assignment) by simple majority (the cutpoint). This design works because at the electoral margin, winners and losers are indistinguishable but for the fact of having obtained office.

Recent scholarship suggests that electoral RDDs are generally valid (1). Caughey and Sekhon (2) discuss in detail the conditions under which margin of victory is a valid forcing variable for RDDs. They explain that margin of victory can be used only when all relevant pretreatment covariates are continuous at the cutpoint (including any unobserved covariates that might confound the treatment effect). They also identify two chief obstacles likely to produce covariate discontinuity (in either observed or unobserved variables) across the cutpoint in electoral RDDs: first, precise manipulation, or the ability of strong candidates (or their parties) to invest just enough resources to achieve a bare majority; and second, fraudulent returns, in which corrupt officials assign bare majorities to their preferred candidates who lost in reality.

We believe neither of these potential obstacles to identification is present in our study. With regard to precise manipulation, Zambian district council elections are conducted in a comparatively low-information environment, making it implausible that politicians possess enough knowledge to engage in the precise manipulation of an electoral margin. Politicians confirmed this point in interviews. With regard to fraudulent returns, Zambia is one of the few countries in Africa in which citizens of all parties have high confidence in electoral institutions, and it has the highest proportion of citizens stating they would fight for democracy if they did not trust the election results (3). Moreover, research on electoral fraud in developing countries generally finds that manipulated elections are won by large margins (4), whereas we are concerned with close elections. Finally, in the 2006 elections we used for our discontinuity design, the ruling party was less successful in council elections than in parliamentary or presidential elections. This suggests that, if the ruling party attempted to manipulate council elections, either the popularity of the opposition party at the local level was *much* greater than expected, or the ruling party's efforts to manipulate local elections were unsuccessful. Taken together, this evidence suggests that close elections at the local level of government in Zambia in 2006 were not fraudulently manipulated.

Empirically, we find that there is covariate balance across the cutpoint of electoral victory. We measure independent covariates through questions on the pretest survey, completed by every participant. The survey contains 59 questions about language group, geographic origin, gender, age, education level, income, business ownership, incumbency, political experience, and party membership. Table S1 uses Kolmogorov-Smirnov tests to show that there is covariate balance across treatment and control groups in the locality of the cutpoint. That is, winners and losers are similar on all covariates except for the treatment of holding office. The only exception is a marginally significant difference in likelihood of owning a car, but with 24 covariates and an alpha of 0.05, the likelihood of observing at least one such significant difference is 71%. A Bonferroni adjustment for multiple comparisons is appropriate in this case (5) and leaves no significant differences. The covariate balance results presented in Table S1 are robust to restricting the sample to those with higher comprehension quiz scores, discussed below, and when restricting the sample to only those individuals who have data for the Trust Game or both the Trust Game and the Dictator Game.

We used six of the covariates in Table S1 to estimate propensity scores for each individual. Those covariates are sex, age, education, income, membership in the ruling party, and incumbency status. Based on the range of estimated propensity scores, $e(X) \in [.23, .82]$, all subjects fall well within the optimal subpopulation for estimating treatment effects (6). This means that there are no subjects whose covariate values essentially guarantee treatment or control. The tight distribution of propensity scores (shown in Figure S1) is further evidence for commensurability of winners and losers at the electoral margin.

One possibly important unobserved covariate could be winners' and losers' differential access to campaign funding. Although we do not have perfect measures of this covariate, in a July 2011 follow-up survey among a subset of participants, 72% of respondents said that all candidates spend about the same amount or that winners of local government elections spend *less* money on campaigns. Given the observed covariate balance and the lack of evidence of electoral fraud or uneven access to campaign finance, we have good theoretical and empirical reason to believe we have a strong RDD that will yield an unbiased estimate of the treatment effect of office holding on behavior.

Figure S1. Histogram of Propensity Scores

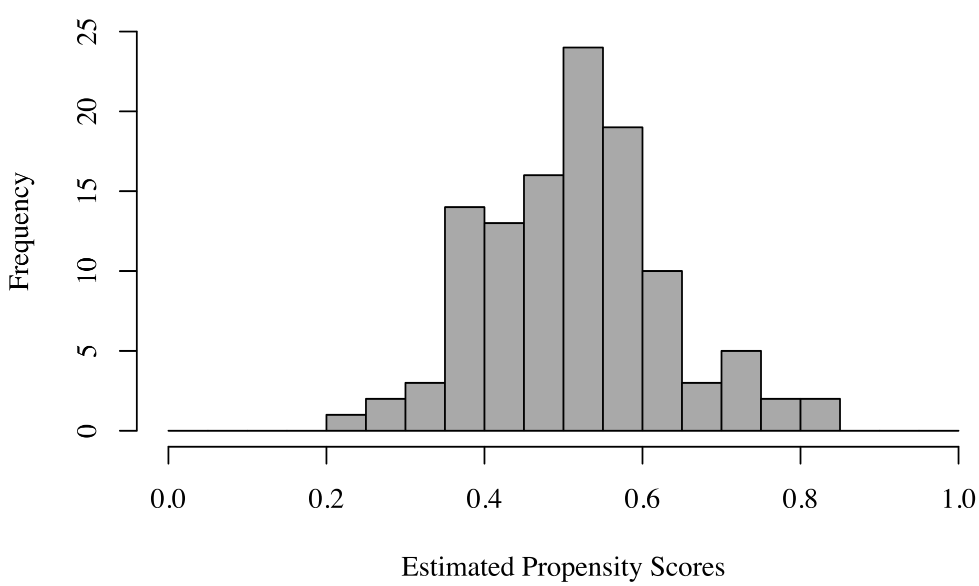


Table 1. Covariate Balance

Covariate	Value	Winners (Treatment)	Losers (Control)	K-S *p*-value
Ethnicity	Bemba	60.8%	59.4%	0.87
Province	Copperbelt	30.2%	25.0%	0.53
Province	Eastern	12.7%	8.9%	0.51
Province	Lusaka	9.5%	8.9%	0.91
Province	Luapula	1.2%	3.4%	0.50
Province	Northern	14.3%	12.5%	0.78
Province	Northwestern	11.1%	12.5%	0.82
Province	Southern	7.9%	14.3%	0.23
Province	Western	1.6%	5.4%	0.27
Sex	Female	10.3%	11.1%	0.87
Age	Age	49.3	50.3	0.54
Education	Education Bracket	5.8	5.9	0.50
Occupation	Farmer	65.8%	62.9%	0.73
Income Bracket	Income Bracket	5.1	4.4	0.12
Ownership	Business	67.1%	67.2%	0.99
Ownership	House	93.5%	92.1%	0.75
Ownership	Car	31.6%	15.9%	0.03*
Ownership	TV	75.3%	65.6%	0.21
Political Experience	Prev. Held Office	14.5%	14.3%	0.98
Political Experience	Family in Politics	53.6%	57.1%	0.77
Incumbency Status	Incumbent	13.9%	14.1%	0.98
Party	MMD	51.3%	39.7%	0.17
Comprehension	Task Quizzes	83.0%	92.0%	0.11
Comprehension	Math Questions	86.0%	89.0%	0.59
Observations		**79**	**64**	

S2. Sampling and Recruitment

Our theoretical population was any politician in Zambia who ran in the 2006 local government elections and won or lost by a margin of less than 10%. There were 1,421 wards with elections in 2006, and 173 (12%) had races ending in a margin of victory of less than 10%, creating a theoretical population of 346 winners and losers. Voter turnout in this election was 71%, and approximately half (750) of the seats were won by the major party, MMD.

Due to the logistics and expense of recruiting politicians in a developing country with low population density and weak communication networks, our sampling procedure was non-random. After obtaining a letter of endorsement from the Local Government Association of Zambia, we first contacted all districts and mailed or faxed letters addressed to the councilors in our theoretical population asking for their assistance in locating candidates in this population. We followed up the letters with telephone calls to the district commissioner or mayor of each province, who then typically gave us councilors' telephone numbers directly. In districts without active offices, fax machines in their offices, or access to reliable postal service, we broadcasted announcements on the radio to recruit subjects. This latter tactic was used in Luapula, Northern, Northwestern and Eastern provinces, where the communication networks were least reliable.

Our recruitment procedure ultimately resulted in telephone contact with 150 of the politicians in our sample. Including all individuals in our theoretical population as potential contacts and restricting the final number contacted to only those with whom we spoke on the telephone, our recruitment procedure yielded a contact rate of 43%. Once we found and spoke with the candidates, our response rate was very high: 95% of the politicians we spoke with and invited to participate ultimately did so.

Our final sample contains 143 politicians from every province and 72% of the districts in Zambia. There are 131 men and 14 women, and the average age is 50. The most common occupations are farmer (65%), business owner (18%), and nonprofit worker (10%). Most of our participants (104, 73%) come from 52 matched pairs of winners and losers (politicians who ran for the same seat), but 39 are unmatched, resulting in 79 winners and 64 losers. Candidates faced between one and four opponents in their elections, with a mean of two. In our sample, 45% of the politicians belong to the MMD party, 48% to opposition parties, and 7% are independents. Fifteen percent of the sample had held another government position before running for councilor, and 16% of the sample had run for councilor before, but 72% of the subjects had never engaged in politics or worked for government before running for councilor in 2006.

The final sample is similar to the broader theoretical population of close winners and losers. A Kolmogorov-Smirnov test reveals that there are no significant differences between the two in the distribution of provinces, vote share, or party affiliation. The Local Government Association of Zambia (LGAZ) assured us that there were no governance or anti-corruption campaigns underway at the time of our research that would affect the behavior of our subjects. In general, the LGAZ stated that the time period of our research and surrounding electoral term (2006-2011) represented a typical experience of holding local government office in Zambia.

S3. Information about the Zambian Context

We recruited Zambian district councilors as the subjects in our games because they make important decisions that affect the daily lives of citizens and involve tradeoffs between their own private welfare and broader public welfare.

A secondary benefit of studying district councilors is that the treatment of holding this particular office is limited and fairly straightforward: a district councilor attends council meetings and exercises legislative powers, but little else in his life changes. Councilors do not receive a new stream of income or need to spend a significant amount of time in the capital city, as do members of the National Assembly. For many elected positions in other contexts, winning office triggers the simultaneous onset of several life changes, each of which might plausibly affect a politician's behavior; the fact that Zambian district councilors' lives change in only a few narrow ways directly related to office holding makes the interpretation of any treatment effect especially clear. At the same time, the officials and politicians we interviewed stated the three mechanisms driving the increase in a new officeholders' level of reciprocity took hold swiftly after taking office and subsequently persisted over time. This leads us to believe that the timing of our experiments (conducted in 2010) relative to the last district elections (held in 2006) is unlikely to drive our findings.

Each district council consists of 8-28 local government officials who control an average budget of around half a million dollars. These are heterogeneous legislative bodies: only one of the district councils (out of 58) consists of members from a single ethnic group, and only five consist of members from a single party. Councilors may spend any and all revenue the district raises locally from 16 different possible sources. The average district allocates funds in six different public-goods categories each year. Zambia has no elected provincial administration, and district officials enjoy extensive spending authority over their local budgets, providing opportunity to make choices that serve either private or public interests (7, 8).

S4. Discussion of Experimental Procedure

Full protocols are available on Mathew D. McCubbins' Dataverse site: https://dataverse.harvard.edu/dataverse/mccubbins.

S4.1. Order of Experimental Tasks and the Potential for Learning

The order of activities in our protocols was not randomized; all subjects experienced the activities in the same order. This could conceivably raise concerns regarding our reciprocity finding, but only if the order of events differentially impacted winners and losers. We do not believe this is a significant confound in our game for three reasons. First, consider the difference between a subject's transfers in the Trust 1 and Dictator tasks. If winners and losers react differently to the order of tasks (transferring either more or less as time goes on), then this difference should vary between these groups. In fact, it does not: $t(92) = -0.71$, $p = 0.48$. Second, winners transferred (significantly) more in Trust Stage 2 but (insignificantly) less in the Dictator Game, so if our reciprocity finding is in fact driven by ordering effects, those effects must be nonlinear. Third, differences in behavior between the Trust Stage 2 and Dictator tasks cannot be driven by learning effects, because subjects learn no new information about their earnings or about the behavior of others between the Trust Stage 2 and Dictator tasks.

S4.2. Anonymity of Partners

Here's what we told our subjects about their partners:

> This experiment consists of many different activities. For some of the activities you will be acting alone. For other activities, you will be randomly matched with another person. Sometimes you will be randomly matched with many people. People you are matched with will always know the exact same information about the activity that you know, except that they will not know what choices you make, and likewise, you will not know what choices they make. No one participating in this experiment will ever know who they are matched with, and as I said before, they will not know the choices made by the people with whom they are matched. Thus, you will not know the choices of the people with whom you are matched and they will not know your choices.

We matched subjects with anonymous partners for three reasons: first, we didn't want to create experimenter demand; second, the existing literature demonstrating the external validity of behavioral games draws from games with anonymous pairing; and third, we weren't solely interested in politicians' behavior toward their peers. For all these reasons, we felt that maintaining the anonymity of subjects' partners would increase our ability to learn from the study theoretically. We turn now to a detailed discussion of these three reasons.

Our first reason for not giving subjects more information about their partners was that we wanted to avoid inducing experimenter demand, and felt that labeling subjects' partners as "a politician" (or any other type of person) would undercut this objective. A nearly universal finding in the experimental literature is that subjects attempt to infer experimenters' interests and objectives (9, 10). We felt that telling subjects “you are matched with another Zambian politician” would introduce a rhetorical treatment that might telegraph experimenter intentions. (As Eckel et al. put it, “The norm is to use neutral language [such as] ‘first mover,’ ‘second mover,’ ‘counterpart,’ to avoid experimenter demand” [11].) Depending on how subjects believed we wanted them to behave toward politicians, this could influence subjects to transfer more or less money (12). Since the effects of non-neutral framing are still poorly understood (13-17), we felt that using neutral frames such as “another person” would clarify the theoretical implications of our work.

Our second reason for using a neutral frame is that the existing experimental literature demonstrating the predictive validity of behavioral games draws on experiments with neutral framing (e.g. 11, 18-28). That is, subjects’ choices in behavioral games are only known to be a valid proxy for real-world behavioral tendencies when experiments employ neutral frames. For this reason, we felt it was important to use neutral frames in our experiments. If we were to deviate from this practice, we couldn’t confidently identify what our experimental tasks were measuring. Had we used the politician frame to describe our subjects’ partners, it’s possible that our subjects’ choices would be less a reflection of underlying behavioral tendencies and more a reflection of their attitudes toward some particular class of people. Since the theory that behavioral games predict real-world behavior is validated by experiments with neutral framing, we chose to use a neutral framing for our protocols. This allows us a tighter connection between the behavioral theory and our predictions regarding politicians’ reciprocity.

Our third reason for choosing a neutral frame rather than labeling partners “politicians” is that we weren’t solely interested in subjects’ behavior toward other politicians. We note in the paper:

> This article focuses on reciprocity. Reciprocity underlies all systems of exchange for mutual advantage. . . . In politics, these reciprocal exchanges can involve office-holders, bureaucrats, non-governmental actors, and citizens. For example, electoral representation, clientelism, patronage, and lobbying are all forms of reciprocity between political elites and citizens. Treaty law, federal grants, party organization, and logrolling are all forms of reciprocity exclusively between elites.

Since we’re interested broadly in whether the experience of holding office makes politicians more inclined to engage in reciprocity, we felt that using neutral framing would allow us to learn more from the study theoretically than if we had specifically primed subjects to think of their partners as politicians. While a larger-scale study could potentially benefit from adding a randomized partner-identity framing treatment, we believe the use of neutral frames clarifies the interpretation our experimental results.

S4.3. Subject Understanding of Instructions

Scholars conducting lab experiments in the field or on a campus often systematically test and report on subject comprehension of the tasks. Not only should subjects’ understanding be tracked and enhanced via tests and subsequent corrections of responses, these data should be used to narrow the pool of subjects. Conclusions about the population’s behavior cannot be drawn if some members of the population are acting under mistaken beliefs about the incentives they face. Chou et al. (29) show that even American college students have difficulty comprehending classic behavioral games; this problem is likely to be enhanced in the low-education environments of developing countries.

As a tool for measuring decision-making, behavioral games can provide very precise and useful measures, but only if subjects can read and perform mathematical operations at the level required to play the game. Without testing and reporting comprehension, it is difficult to assess the usefulness of studies employing behavioral games, even if findings are reported to be statistically significant. For this reason, our protocols included two main methods for testing our subjects' comprehension of the protocols: a survey administered before the experiment, and quizzes administered during the experiment.

The pen-and-paper pretest survey was itself inherently a measure of subjects' ability to read and respond to questions in their home language, but it also included questions gauging basic arithmetic skills. In the experimental session, an enumerator read instructions to a group of subjects, then handed out an incentivized pen-and-paper quiz about the instructions to each subject. Participants received \$0.60 for each correct answer and our enumerators reminded them of this incentive immediately before each quiz. Subjects had a pre-specified period of time to answer the questions, after which enumerators reviewed each respondent's answers, correcting incorrect answers with a red pen. The average participant score on these quizzes was 87%.

We used the data from the surveys and comprehension quizzes to restrict our sample to include only those subjects who (A) were able to demonstrate the mathematical skills to participate without guessing, and (B) comprehended the game instructions. Although we believe our task-specific comprehension tests enabled subjects to learn from the quizzes and refine their knowledge accordingly, we do not want to assume learning happened without testing for it. We use the percentage of questions participants got correct to determine the whether they are included in our sample. We removed 18 participants with low quiz scores, leaving a sample of 125. Of the 18 excluded from analysis, 13 were winners (17% of the sampled winners) and 5 were runners up (8% of the sampled runners up), providing evidence that playing these games is not a skill obtained in office, and we are therefore not conditioning on a post-treatment variable by excluding those that could not demonstrate comprehension. More importantly, overall there was no statistically significant difference in means between the two groups on their comprehension score (p-value = 0.12).

S4.4. Ecological Validity

We asked the members of our sample to participate in a day of experiments in Lusaka, Zambia's capital. We went to great lengths to run the experiments in an environment that resembles as closely as possible the environment in which our subjects make their political decisions, and set the stakes high enough to ensure that our subjects made their decisions in our study carefully. One benefit of using politicians as subjects is that politicians have experience in making decisions in formal environments. Local political leaders are regularly brought to workshops (generally in hotel conference rooms) where they are asked to help make decisions and are paid for their participation. We designed our experiments to be as similar as possible to the format of those workshops.

To avoid influencing subjects' behavior, we never interacted with subjects nor were we visible to them. Instead, we trained the same local research assistants who translated our protocols and survey to conduct the experiments in order to avoid any confounding effects that might arise from the presence of foreign researchers. We used a variety of methods to minimize experimenter demand. We repeatedly reminded the subjects that their answers were confidential and anonymous. We assigned each subject a random identification number and tracked their actions only using this number. None of the experimenters in the rooms with subjects ever got to observe subjects' choices. We crafted the protocols and their translations to avoid suggestive linguistic framing. We also reminded them repeatedly of the payoffs in the games—payoffs that made the opportunity cost of fulfilling any perceived experimenter demand instead of maximizing their own welfare very high.

We first tested our experimental protocols with undergraduate students at a large, Southern Californian university, at the University of Zambia, and with a small sample of politicians outside our theoretical population. We asked for their feedback on the protocols—especially the clarity and appropriateness of our questions—in post-experiment surveys and focus-group discussions. Based on feedback from these test subjects and from our team of translators, we revised the protocols to clarify the language, simplify some of the decisions, and avoid language that could generate experimenter demand. We also chose to provide a calculator at every desk, as we observed in the pilot test that politicians frequently used calculators on their mobile phones for mathematics.

The experiments were carried out using paper and pencil. We chose not to use computers because local government politicians in Zambia are not frequent computer users. Each experimental session had approximately 20 subjects. For each task, we randomly matched subjects with someone in another location and reminded them of this set up every time they made a decision. We made this distinction to prevent their behavior from being affected by friend, familial or tribal ties of someone in their room, and to mimic more closely the heterogeneous environments in which they make political decisions. A roughly equal number of winners and losers played games in each room and each session.

The maximum a player could earn in the Trust Game is 20,000 Zambian Kwacha (US$4.20), and the minimum is 0 Kwacha. Given that most of the Zambian subjects self-report earning between K250,000-K500,000 ($52.50-$105) per month, the difference between maximum and minimum payoffs is more than one day's income, and for some participants, more than two days' income. We believe we incentivized the experiments at a high enough level to motivate politicians to devote serious thought before making choices. Since the electoral winners and losers in our sample had the same average self-reported income, the value of the incentives was on average the same for winners and losers. Since the final amount of payment was contingent on subject choices, we are confident that politicians considered each case carefully.

Although the payments incentivized subjects to think judiciously, we did not prime them to think about their normative obligations to constituents or frame these choices as moral ones with right or wrong answers. Our instructions made it very clear that individuals would be paid for their actions, and it was implied that they would be financially compensated for making decisions according to the rules of the activities and what they thought prudent, not for making decisions according to other behavioral norms. Our translators and pilot tests with Zambian students helped us to describe rules with language as free as possible of words that might prime certain actions. In advance of participating in our experiments, the subjects received a letter stating that the project was interested in studying how local leaders make decisions.

The experiments were conducted in July and August of 2010. This was in the middle of the local government term and the fiscal year. Campaigning had not yet started for the 2011 elections, and the Local Government Association of Zambia confirmed that no recent anti-corruption campaigns or trainings for local government officials had occurred.

S5. Robustness Tests and Alternative Specifications

Our sampling design was to recruit only from those close to the cutoff and our analytical approach was to determine the model specification before we entered the data. Both of these choices guard against manipulating the analysis.

We believe the appropriate specification is the one we adopted before we ever began the analysis: $Y = \beta_0 + \beta_1 V + \beta_2 T + \varepsilon$. We add an interaction term as a robustness test because some scholars prefer that specification (1). Gelman and Imbens argue that it is not appropriate to include higher-order polynomial functions of the forcing variable (30), but in the interest of transparency we include quadratic and cubic functions of the forcing variable in Table S2.

We note that the main result holds in the robustness check (linear-interaction) and in one of the alternative specifications (quadratic). The result does not hold under quadratic-interaction or higher-order polynomial specifications, but then we do not have any expectation that it should. As we note in the main text, the parsimonious linear model aligns best with our lack of theory to explain a non-linear relationship between margin of victory and reciprocity (or any other behavioral traits).

In Table S3 we include those with low comprehension scores in our analysis (see Section S4.3). The results are essentially unchanged, though the effect size is smaller, which is consistent with our expectation that those who did not comprehend the tasks are merely adding random noise to the data.

Finally, we make one additional specification note: the models presented in the manuscript do not include heteroskedasticity-robust standard errors. When we include these, the results are similar but the p-value increases to 0.0647.

Table S2. Effect of Office on Reciprocity: Robustness Test and Alternative Specifications

	Original Model	**Robustness Test**	**Alternative Specifications**			
Specification	**Linear**	**Linear-Interaction**	**Quadratic**	**Quadratic-Interaction**	**Cubic**	**Cubic-Interaction**
Electoral Victory	.356** (.171)	.335* (.171)	.349** (.172)	.0610 (.251)	.160 (.229)	.206 (.341)
Margin	-2.21 (1.53)	.355 (2.21)	-2.17 (1.53)	20.2** (8.37)	2.01 (3.71)	11.2 (22.0)
$Margin^2$			-14.4 (15.0)	195** (79.4)	-16.8 (15.1)	-29.4 (510)
$Margin^3$					-477 (385)	-1,490 (3,330)
Victory* Margin		-4.84 (3.04)		-30.3*** (11.5)		-29.5 (29.8)
Victory* $Margin^2$				-137 (112)		303 (701)
Victory* $Margin^3$						10.2 (4,660)
Constant	-.0809 (.0979)	.047 (.126)	-.032 (.110)	.407** (.191)	.071 (.138)	.328 (.263)
Observations	95	95	95	95	95	95
R^2	.050	.076	.059	.139	.075	.143
Adjusted R^2	.029	.045	.028	.091	.034	.074

*Note: *p<0.1; **p<0.05; ***p<0.01*

Table S3: Effect of Office on Reciprocity: With and Without Non-Compliers

	Original Model (Without Non-Compliers)	**Alternative Model (With Non-Compliers)**
Electoral Victory	.356** (.171)	0.287* (.166)
Margin	-2.21 (1.53)	-1.32 (1.46)
Constant	-.0809 (.0979)	-0.0494 (.0945)
Observations	95	106
R^2	.050	0.041
Adjusted R2	.029	.023

*Note: *p<0.1; **p<0.05; ***p<0.01*

S6. Choice of Bandwidth for Regression Discontinuity

Our RDD is very unusual. Most RDDs rely on observational data to measure the dependent variable. For this reason, the analyst usually begins with the full set of data and must select (using tools like the *rdrobust* package in R) an appropriate bandwidth within which assignment to treatment is (conditional on the forcing variable) "as-if random."

We did not have the luxury of using extant data to measure our dependent variable, reciprocity; instead, we had to recruit subjects to a conference center where we could use an experiment to generate behavioral data from which we calculate Reciprocity Scores. This procedure was costly, so instead of recruiting politicians from across the full spectrum of the forcing variable, we only recruited politicians within 10% of the electoral cutpoint. Thus we pre-specified our bandwidth before data collection began and only collected outcome data within this bandwidth. This essentially "hard-coded" our chosen bandwidth for analysis, preventing us from manipulating the bandwidth to obtain significant results.

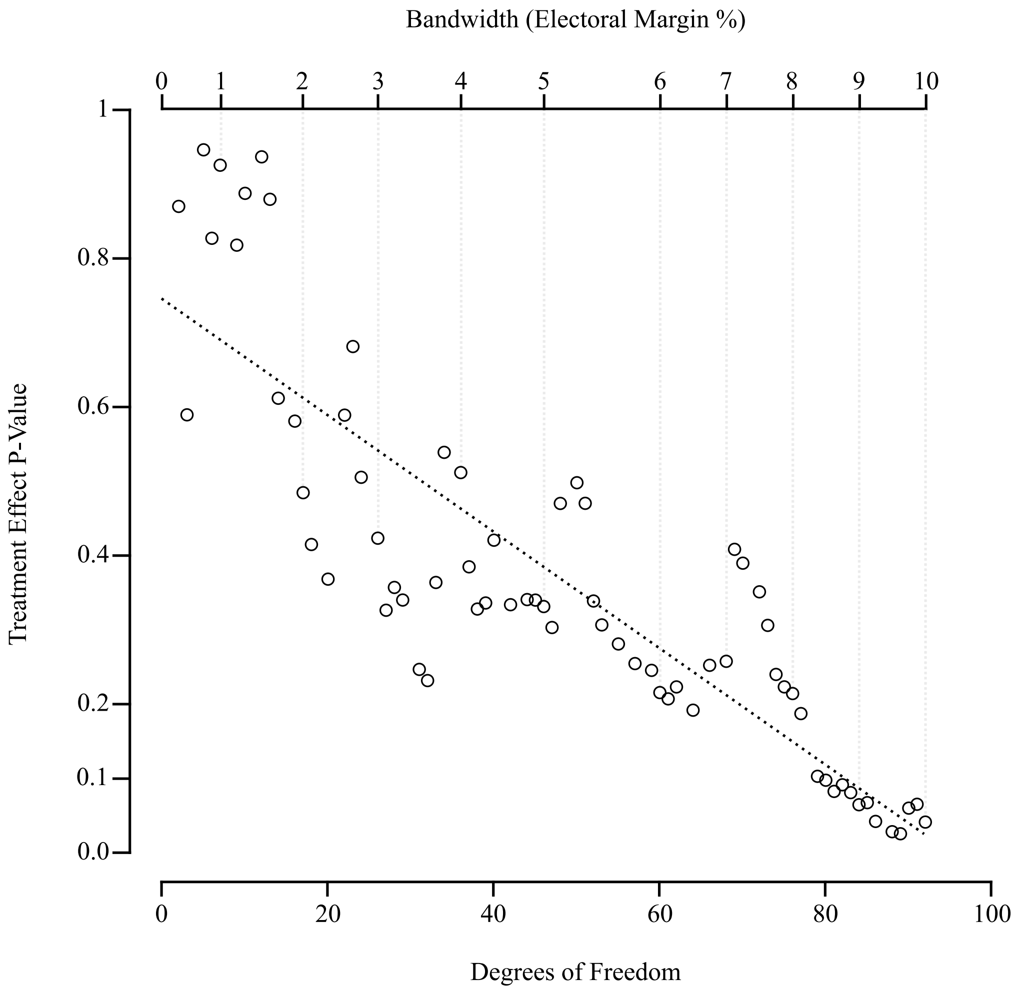
As a result, there are no real bandwidth choices to be made, because (A) we only collected data from individuals close to the cutoff, and (B) it was not possible to recruit enough individuals to tighten the bandwidth without losing statistical power. The first point is a virtue of our design, because by hard-coding the bandwidth into the sampling procedure, we guard against post-hoc manipulation of bandwidth (p-hacking for results). The second point is a necessity of our design, because we picked the smallest bandwidth that would allow us to assemble a sample of politicians large enough for causal inference.

There were only 346 individuals who won or lost local government elections in Zambia in 2006 by a margin of less than 10%. That we were able to recruit 143 politicians (41%) to participate in our research sessions in Lusaka is somewhat of a miracle. We used every means at our disposal—announcements at the Local Government Association of Zambia (LGAZ) annual meeting, recruitment letters disseminated through LGAZ to the district offices, constituency-specific radio ads asking specific councilors to call our call center, recruiting bonuses, and an incredibly diligent team of research assistants calling, emailing, and visiting government officials across the country—to recruit as many of these individuals as possible. Our final sample contains politicians from every province in the country and from 71% of the districts. If we could have recruited more politicians to participate, we would have. Unfortunately, however, we did not end up with a sample size large enough to allow us to reduce the bandwidth below the 10% we initially chose.

In Section S1, we provide empirical evidence that our choice of bandwidth was valid. Because of the excellent covariate balance between election winners and losers, and the fact that we hard-coded bandwidth for the regression discontinuity into our sampling procedure, we do not feel that post-hoc bandwidth robustness checks are necessary.

For the purpose of transparency, we ran a regression for every possible bandwidth, by removing subjects one at a time in decreasing order of distance from the cutpoint. Figure S2 shows the degrees of freedom and treatment-effect *p*-value for each of these regressions. All regressions with a bandwidth of 8.8% or greater result in treatment-effect estimates of $p < 0.1$. We also regressed the p-values from these regressions on their degrees of freedom (the black dotted line in Figure S2). The strongly linear relationship between degrees of freedom and p-value confirms that as the bandwidth narrows, the shrinking sample size drives up uncertainty.

Figure S2. Increasing *p*-value of treatment effect as bandwidth narrows

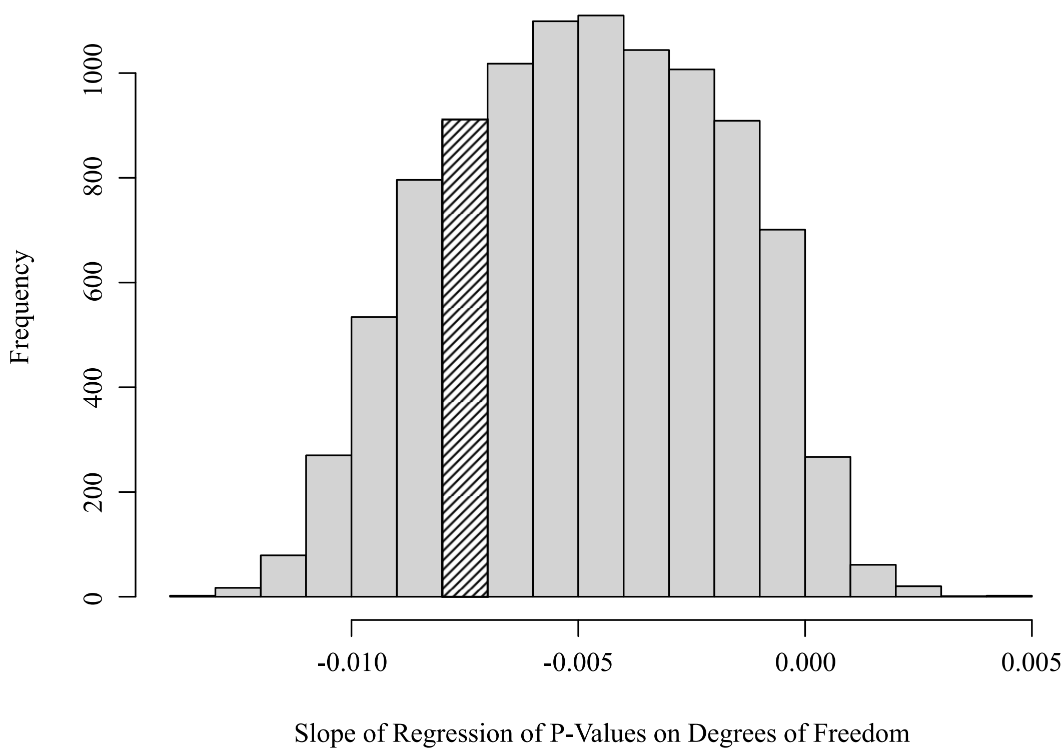


After running every possible regression by iterated removal of the largest-margin case, we wanted to confirm that the relationship between treatment-effect p-values and degrees of freedom was driven merely by shrinking sample size rather than the changing bandwidth. If the effect we report on in the paper were sensitive to bandwidth selection, this would imply that the data farthest from the cutpoint was essential to our result—and thus that the relationship between p-values and degrees of freedom would be very highly negative, since the removal of cases far from the outpoint would produce a precipitous drop in p-values.

In order to show that the observed relationship between p-values and degrees of freedom is not driven by the reduction in bandwidth, we have to compare it to the expected relationship given a reduction in sample size *but not* bandwidth. To do this, we generated 10,000 random orderings of the cases, and then for each ordering ran regressions on every possible sample size by removing cases in that random order, resulting in nearly 1,000,000 regressions. For each set of regressions generated by one of the random orderings, we regressed p-values on degrees of freedom, giving us 10,000 coefficients describing the slope between treatment-effect p-values and degrees of freedom.

Figure S3 shows the distribution of those coefficients. The slope from our bandwidth-ordered sample-size reduction (-.0078) falls in the crosshatched bin of coefficients, close enough to the mean (almost exactly one standard deviation) to give us confidence that any failure to identify significant treatment effects using a narrower bandwidth are driven by lack of statistical power.

Figure S3. Distribution of coefficients relating treatment *p*-values to degrees of freedom across 10,000 random-order iterated removals of cases. The coefficient for bandwidth-ordered removal of cases is -.0078 (the crosshatched bin).



S7. Alternative Dependent Variables

We next include a different operationalization of reciprocity as the outcome variable. Specifically, rather than subtracting out the Dictator Score from the Trust Player 2 choice, we divide the Trust Player 2 choice by the endowment and use this as the outcome variable. This provides a measure of "trustworthiness." Trustworthiness is related to reciprocity in that it captures to what extent the subject will return a favor done unto them by their partner, but it includes the generosity component that the Reciprocity Score subtracts out (18). Trustworthiness, a reaction-based trait, is distinct from Trust, a first mover trait. The results on Trustworthiness (Table S4) are consistent with the finding in the main text.

We're grateful to an anonymous reviewer for suggesting the next robustness test, as this approach enables us to retain a larger sample of subjects for our analysis, even if we have missing data for them for either the Trust Game or Dictator Game. Here, we stack the data from the Dictator Game and Trust Game and use a variable representing the share of the endowment transferred as the dependent variable. This result of this analysis appears in Table S5. This robustness check aligns with the results in the main text. Controlling for endowment size, in-office officials in the second stage of a Trust Game transfer a significantly higher percentage of their endowment than out-of-office officials. This difference does not hold for the Dictator Game, suggesting that in-office officials are conditioned by the treatment of holding office to respond to gestures of good faith, or in other words, to adhere to the norm of reciprocity. They are not, however, conditioned to be more generally generous (or selfish), which is the trait that the Dictator Game is often thought to measure. Further, losers appear to be less sensitive to the different games, behaving consistently across the Trust Game and Dictator Game. Finally, the endowment level does not appear to significantly predict the share transferred.

Table S4: Trustworthiness

	Trustworthiness
Winner	0.327** (0.128)
Margin	-2.393** (1.132)
Constant	0.149** (0.071)
Observations	104
R^2	0.062
Adjusted R^2	0.043

*Note: *p<0.1; **p<0.05; ***p<0.01*

Table S5: Stacked Games

	Trust 2 Transfer
Winner	-0.046 (0.055)
Trust Game	0.025 (0.064)
Endowment	0.002 (0.004)
Winner*Trust	0.156** (0.080)
Constant	0.212*** (0.062)

*Note: *p<0.1; **p<0.05; ***p<0.01*

S8. References

1. Eggers AC, Fowler A, Hainmueller J, Hall AB, Snyder JM (2015) On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races. *Am J Pol Sci* 59(1):259-274.
2. Caughey DM, Sekhon JS (2011) Elections and regression discontinuity design: lessons from close U.S. house races, 1942-2008. *Polit Anal* 19(4):385-408.
3. Moehler DC (2005) Free and fair or fraudulent and forged: elections and legitimacy in Africa. *Afrobarometer Working Paper No. 55.*
4. Simper A (2008) Cheating big: on the logic of electoral corruption in developing countries. *University of Chicago.*
5. Lee W-S (2011) Propensity score matching and variations on the balancing test. *Empir Econ* 44(1):47-80.
6. Crump RK, Hotz JV, Imbens GW, Mitnik OA (2009) Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96(1):187-199.
7. Das J, Dercon S, Habyarimana J, Krishnan P (2003) Rules vs. discretion: public and private funding in Zambian basic education. Part I: funding equity. *The World Bank Working Paper Series No.* 62.
8. Gibson CC, Hoffman BD (2013) Coalitions not conflicts: ethnicity, political institutions, and expenditure in Africa. *Comp Polit* 45(3):273-290.
9. Orne MT (1962) On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *Am Psychol* 17(11): 776-783.
10. Nichols AL, Maner JK (2008) The good-subject effect: investigating participant demand characteristics. *J Gen Psychol* 135(2):151-166.
11. Eckel C, De Oliveira A, Grossman PJ (2008) Gender and negotiation in the small: are women (perceived to be) more cooperative than men? *Negotiation Journal* 24(4): 429-445.
12. McCubbins MD, Turner M (2012) Going cognitive: tools for rebuilding the social sciences. *Grounding Social Sciences in Cognitive Sciences* 387.
13. Abbink K, Hennig-Schmidt H (2006) Neutral versus loaded instructions in a bribery experiment. *Experimental Economics* 9(2):103-121.
14. Barr A, Serra D (2009) The effects of externalities and framing on bribery in a petty corruption experiment. *Experimental Economics* 12(4):488-503.
15. Baldry JC (1986) Tax evasion is not a gamble: A report on two experiments. *Econ Lett* 22(4):333-335.
16. Alm J, McClelland GH, Schulze WD (1992) Why do people pay taxes? *J Public Econ* 48(1):21-38.
17. Zizzo DJ (2010) Experimenter demand effects in economic experiments. *Experimental Economics* 13(1):75-98.
18. Cox JC (2004) How to identify trust and reciprocity. *Games Econ Behav* 46(2):260-281.
19. Glaeser EL, Laibson DI, Scheinkman JA, Soutter CL (2000) Measuring trust. *Q J Econ* 115(3):811-846.
20. Uslaner EM (2002) The moral foundations of trust. (Cambridge University Press).

21. Guerra G, Zizzo DJ (2004) Trust responsiveness and beliefs. *J Econ Behav Organ* 55(1): 25-30.
22. Karlan DS (2005) Using experimental economics to measure social capital and predict financial decisions. *Am Econ Rev* 95(5):1688-1699.
23. Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E (2005) Oxytocin increases trust in humans. *Nature* 435(7042):673-676.
24. Bellemare C, Kröger S (2007) On representative social capital. *Eur Econ Rev* 51(1): 183-202.
25. Sutter M, Kocher MG (2007) Trust and trustworthiness across different age groups. *Games Econ Behav* 59(2):364-382.
26. Johnson, ND, Mislin A (2008) Cultures of kindness: A meta-analysis of trust game experiments. Available at Social Science Research Network: http://ssrn com/abstract, 1315325.
27. Croson R, Gneezy U (2009) Gender differences in preferences. *J Econ Lit* 47(2):448-474.
28. Van Den Bos W, van Dijk E, Westenberg M, Rombouts SA, Crone, EA (2009) What motivates repayment? Neural correlates of reciprocity in the Trust Game. *Soc Cogn Affect Neurosci* nsp009.
29. Chou E, McConnell M, Nagel R, Plott C (2009) The control of game form recognition in experiments: understanding dominant strategy failures in a simple two person 'guessing' game. *Experimental Economics* 12(2):159-179.
30. Gelman A, Imbens G (2014) Why high-order polynomials should not be used in regression discontinuity designs. *NBER Working Paper Series* 20405.