Journal-Based Replication of Experiments: An Application to "Being Chosen to Lead"*

Allan Drazen
University of Maryland
drazen@econ.umd.edu
econweb.umd.edu/∼drazen/

Anna Dreber
Stockholm School of Economics
anna.dreber@hhs.se
sites.google.com/site/annadreber/

Erkut Y. Ozbay
University of Maryland
ozbay@umd.edu
econweb.umd.edu/∼ozbay/

Erik Snowberg
University of British Columbia, CESifo, NBER
snowberg@mail.ubc.ca
eriksnowberg.com

May 15, 2021

Abstract

Recent large-scale replications of social science experiments provide important information on the reliability of experimental research. Unfortunately, there exist no mechanisms to ensure replications are done. We propose such a mechanism: journal-based replication, in which the publishing journal insists on a replication attempt between acceptance and publication. We discuss what we learned from a proof-of-concept journal-based replication at the *Journal of Public Economics.* Our experience indicates that journal-based replication would be relatively straightforward to implement for laboratory experiments.

JEL Classifications: A11, A14, C18, C92
Keywords: replication, reliability, experiments, journal-based replication

*The authors gratefully acknowledge the support of *Journal of Public Economics* editors, Erzo F.P. Luttmer, Wojciech Kopczuk, and Henrik Kleven for support and advice during the process of replication. We thank Andrew Proctor, Andreas Born, and Edoardo Bollati for excellent research assistance. We thank Magnus Johannessson, the editor Keith M. Marzilli Ericson, and two anonymous reviewers for their extremely useful comments and suggestions. We further acknowledge the generous support of the Canada Excellence Research Chairs program, which provided funding for this study. Dreber thanks the Jan Wallander and Tom Hedelius Foundation, the Knut and Alice Wallenberg Foundation, the Marianne and Marcus Wallenberg Foundation, and the Austrian Science Fund (FWF, SFB F63) for financial support. Drazen thanks the U.S. National Science Foundation (SES 1534132) for financial support.

1 Introduction

An advantage of experimental methods is that they can be *replicated*—that is, the same experiment can be run again and the result compared to the original study (Coffman and Niederle, 2015; Maniadis et al., 2015, 2017; Camerer et al., 2019; Nosek and Errington, 2020).[1] Despite the fact that all scholars benefit, at least weakly, from knowing whether or not a particular result holds in a replicated experiment, the incentives for any one scholar to attempt to replicate an experiment are low, leading to few attempts (Berry et al., 2017; Maniadis et al., 2017). That is, replication attempts are public goods, and like most public goods, are under-provided.

We suggest a mechanism to increase replication attempts—journal-based replication—in which the journal that publishes a study takes responsibility for ensuring a replication of that study is attempted. The idea behind journal-based replication is to focus the responsibility for providing a public good on the parties—journals and authors—who benefit most from it (Olson, 1965). There are many ways to accomplish this, and we develop a framework for choosing policies for journal-based replication in Section 2. We describe some of those policies in Section 3, and based in part on those policies, attempt a proof-of-concept journal-based replication in Section 4. Interestingly, our replication attempt produced the opposite result from the original study—a negative, statistically significant treatment effect rather than a positive one. Based on this fact, and our experience with the proof-of-concept, we add additional nuance to our suggested policies in Section 5. We conclude with a summation of the merits of, and potential issues with, journal-based replication, in Section 6.

Our proof-of-concept occurred at the *Journal of Public Economics*, considered a top field journal in economics. The experimental study selected for replication, with the support of the authors (who became co-authors of the current manuscript), was "Does 'Being Chosen to Lead' Induce Non-selfish Behavior? Experimental Evidence on Reciprocity," by Drazen and Ozbay (2019). That study found that representatives are more responsive to the concerns of their constituents if they were elected rather than appointed, all else equal. This original result was found in a student sample at the University of Maryland, and our replication attempt was performed by a contract lab at the University of Valencia, in Spain. As noted

[1] Replication can refer to both a re-run of the same experiment, or the fact that that re-run produced a similar result. We generally refer to the former as a *replication attempt*, although we use the term replication when the context makes the meaning clear.

above, in the Spanish sample, we found the opposite result—namely that appointed representatives are more responsive than elected representatives, all else equal. We note in Section 4 that this result is consistent with other data showing Spanish participants exhibit different patterns of reciprocity than participants from other OECD countries, although other factors may be at play as well.

2 A Framework for Journal-Based Replication

Recently, concerns about the replicatability of experimental studies have become more pronounced. In psychology, a finding that out of 97 replication attempts of non-null results, only 35 (36%) of them produced similar results to the corresponding original study contributed to a general perception of a "replication crisis" (Open Science Collaboration, 2015). A similar study of experimental economics results found that of 18 studies published in two "top-5" journals 11 (61%) produced similar results (Camerer et al., 2016).[2] This lead to a more muted response, possibly due to the positive comparison with psychology, and that these concerns were folded into a larger discussion about external validity and optimal study design (Kasy, 2016; Banerjee et al., 2017; Christensen and Miguel, 2018; Banerjee et al., 2020).[3] In both economics and psychology, replication attempts have come to be seen as one of several tools, including *pre-analysis plans*—specifying which statistical hypotheses will be investigated before an experiment is run—for ensuring the generalizability of experimental results.

The ability to replicate studies is an advantage of the experimental method, and replication attempts may be superior to other methods that seek to ensure the reliability of experimental results. Observational studies, for example, cannot be replicated, only *reproduced*—that is, a researcher can verify that the same data and code produces the same results, but

[2]These top-line statistics are not straightforward to compare given the differences in sample sizes and inclusion criteria. For example, in the economics study, interaction effects were not included, whereas in the psychology study interaction effects were included and found to be less likely to replicate than main effects. Additionally, the psychology study replicated three null results, although these were not included in the top-line statistics.

[3]For example, there is an ongoing debate about the *credibility* of economics research—that is, whether studies are identifying causal pathways, and whether studies of causal mechanisms produce better estimates of quantities of policy interest (Imbens, 2010; Deaton, 2010). Of particular note in this literature is Frankel and Kasy (2018), which proposes a framework for identifying studies that would contribute the highest marginal value to learning if they were being published. We abstract from this question and assume that any study worth publishing is also worth attempting to replicate.

in most cases cannot re-run a policy experiment in a different population at a different point in time. Coffman and Niederle (2015) argue that the possibility of cheap replications obviate the need for other tools, such as pre-analysis plans. Their argument is that pre-analysis plans prevent researchers from exploring their data and finding new and unexpected phenomena, and provide little protection against false positives. On the other hand, if there is concern that a published study is a false positive, a replication attempt will reveal that fact with a high degree of certainty. This argument requires that replications actually be attempted, which is, unfortunately, rarely the case.

Replication attempts are public goods, and like most public goods, tend to be under-provided. Almost all researchers would (weakly) benefit from knowing if a given study replicated (except for, perhaps, the authors of the original study), but for almost all researchers, the benefit of knowing this is less than the cost of performing a replication. In psychology, large consortia have arisen to replicate prominent experimental studies (Open Science Collaboration, 2015; Ebersole et al., 2016; Klein et al., 2018). Although these efforts should be lauded, they cannot cover all published experiments, and often result in a delay of several years before scholars know if a study replicates. This delay comes at a cost: recent research suggests that results that fail to replicate are cited equally or more often than those that do (Gneezy and Serra-Garcia, 2020; Schafmeister, 2020; Yang et al., 2020). Large-scale replication efforts are much less common in economics: a review of studies published in one volume of the *American Economic Review* concluded that, "a majority of very well-published papers in economics are not being replicated at all" (Berry et al., 2017). Further, a meta-study of replication in experimental economics suggests that replications have only been attempted for 4% of published studies (Maniadis et al., 2017).

Journal-based replication seeks to ameliorate the public goods problem by pushing those who benefit most from replication attempts—authors and journals—to bear the cost of providing them. The simplest possible definition of journal-based replication is that a journal ensures (somehow) that replications of studies published in their journal are attempted. This will likely enlist, in some way, the authors of the original study. As Olson (1965) pointed out, an individual or small group benefiting enough from a public good to bear the entire cost of providing it is one of the primary ways collective action problems are solved. Journals benefit as replications enhance the journal's reputation for transparency and the reliability of studies it publishes. If a study replicates, it will increase the prominence of the article,

benefitting both the journal and the authors. To put this another way, if one wanted to provide a subsidy to anyone wishing to make a replication attempt, the cheapest entities to subsidize should be the authors of the study and the journal that published it.

Implementing journal-based replication is complicated by many practical factors, including the fact that the interests of journals and authors are not always aligned. For example, a journal would prefer to publish studies that it knows replicate, whereas authors would prefer if their studies were published whether or not they replicate. As such, journals would prefer to accept papers after a replication attempt, whereas authors would prefer that their study be accepted for publication before a replication attempt is made.

Some of these practical considerations are eased by recognizing that journals also wish to maximize submissions, and in order to do so, will want to minimize the impact of journal-based replication on author surplus. While several factors feed into this goal—for instance, more submissions likely means that a journal can be more selective, thus increasing the quality of published research—for our purposes it is sufficient to note that these factors result in journals acting as if maximizing submissions is a primary goal in and of itself. Thus, in the example of the prior paragraph, it likely makes more sense to use authors' preferred policy—(conditional) acceptance before a replication attempt—than the journal's. Failing to use this structure might cause authors to favor other outlets for submission.

The rest of the paper is dedicated to trying to work out solutions to the practical questions raised by journal-based replication. In the next section, we use the principles described above—focusing on journals and authors to provide the public good of replication attempts, and acting to ensure that authors receive as much surplus as practical—to create recommendations for dealing with these questions. Some cannot be answered based on these principles alone, so we also bring in perspectives from the literature on replication. We then discuss the results of our pilot of journal-based replication, before exploring what we learned from it, and which questions remain open.

2.1 Alternatives

Before turning to a discussion of practical questions, we first describe three related approaches to improving the reliability of experimental social science research.

Registered reports are essentially journal-based incentive to use pre-analysis plans. In particular, authors submit an experimental design and a pre-analysis plan to a journal for

consideration. If the plan—after modifications suggested by editors and reviewers—is of sufficiently high quality, the journal will then accept it. That is, the write-up of the experimental results—conditional on adhering to the plan—is accepted for publication no matter what the results of the experiment turn out to be; thus, this method also helps combat publication bias (Nosek and Lakens, 2014). Following the logic of Coffman and Niederle (2015) discussed above, this would be a more attractive option than journal-based replication when an experiment, and hence a replication attempt, is particularly expensive. However, few journals have implemented this approach—the *Journal of Development Economics* is the lone economics journal, at this point, with a permanent track for registered reports—perhaps because journals prefer to make the decision to publish an article when they have a better idea of what the results of an experiment are (Coffman and Niederle, 2015; Brodeur et al., 2016).

Another alternative consists of one set of authors writing a working paper with the results of a preliminary experiment, and inviting other authors to join them in replicating the experiment and submitting the paper to journals. We are aware of a single instance of this approach (Butera et al., 2020). Two scholars ran an experiment (Butera and List) which produced an unexpected finding. They wrote up these results as a working paper (Butera and List, 2017), and invited other authors to propose a replication attempt of the original experiment. The replication attempt would then be written up with the original experiment as a joint study by all involved (Butera et al., 2020). As one of the authors of the original study is quite prominent, this is essentially a trade: those conducting the replication attempt get to write a potentially interesting article with a very prominent researcher, and in return, the prominent researcher gets some validation that their original result was not just a false positive. While this is an interesting model, it is not clear if most authors would have significant stature to engage in such a trade.

Finally, many journals have implemented code and data archives for empirical studies to make sure that those studies can be reproduced. The similarity with journal-based replication is that journals have implemented a policy that encourages, often at some costs to the journal, good research practices. Such an archive is complementary to journal-based replication: the code and data for both the original experiment and the replication study should be made available to future researchers.

3 Journal-Based Replication in Practice

We now turn to applying the general principles above—focusing on journals and authors to provide the public good of replication attempts, and acting to ensure that authors receive as much surplus as practical—to answer practical questions of how journal-based replication should operate. When these principles are insufficient we turn to perspectives from the literature on replication. What we believe are the general best practices sometimes diverge from what we implemented in our pilot; this was due to the *Journal of Public Economics* being worried about reputational damage that might occur if the authors had a bad experience with the pilot. Thus, our choices in the pilot can be seen as an extreme version of trying to ensure that authors receive as much surplus as possible, to the neglect of all other considerations. For example, in order to address concerns of the editors of the *Journal of Public Economics* about potential negative reputational concerns due to the pilot, we agreed that the authors of the original study would have control over what was done with the replication results. While we do not view this as practical in general, these choices may still be informative of how far one might push the policies for journal-based replication.

3.1 When Should Replication be Attempted? What Should be Done with the Results?

Conditional on attempting a replication, there are effectively three points when a replication could occur: before acceptance of an experimental manuscript, between (conditional) acceptance and publication, and after publication.

A journal and authors will have conflicting preferences over whether a replication attempt should occur before or after acceptance; with the principle minimizing the impacts on authors implying it should happen after (conditional) acceptance. A journal would like to wait to see if an experiment replicates before accepting for publication so as not to use space for a "null" result, as we assume that editors seek to maximize the impact of the published papers. Authors, on the other hand, would like acceptance to be assured before attempting the replication. In addition to likely decreasing submissions, asking for replications to be completed before acceptance would have the additional downside of the replication data not seeing the light of day, as authors would be free to submit the original article to other journals without such a policy. This, and the principle of trying to direct surplus to authors,

implies that the replication attempt should occur after acceptance.

Should the replication attempt happen before or after publication? Here, the principle of focusing the costs of replication on those who benefit more implies it should occur before publication. Otherwise, a journal may choose not to dedicate space to the replication attempt after the original article has been published, or may not want to diminish the influence of a successful article it has published. As such, we believe the replication attempt should occur before publication, and the results should be integrated into the published paper.

Although journals may be tempted to revoke acceptance after a failed replication, we believe the reputational costs of doing so would be so high that this would not occur in practice.[4] A more plausible cause for concern is that if too many experiments that were accepted for publication failed to replicate, a journal may wish to end their policy of journal-based replication altogether. If this occurred, we would hope that the journal might examine its editorial decision-making process to try to understand why so many experiments that do not replicate are making it through, but such reflection is by no means assured.[5]

Note that accepting the paper before replication is attempted does not mean that the paper should be published "as is." The editor in charge of a paper that has undergone journal-based replication can insist on revisions in light of the replication results, which might mean writing the paper primarily around the replication, rather than around the original results (which would still need to be discussed, but perhaps would received fairly little attention). Thus, the acceptance in this case is much more like a conditional acceptance in a standard editorial process—the paper is accepted conditional on the replication attempt taking place, and suitable changes being made in light of the results of the replication attempt. As our pilot further informed our opinion about what should be done with the results of the replication attempt, we return to this point in Section 5.2.

This was not the path we took in our pilot. As noted above, the pilot was conducted under guidelines that gave authors of the original study have final say over each part of the process. Although the replication attempt was made before publication, the original paper was published with only a footnote mentioning the results of the replication, so as not to

[4] An exception would be if the replication attempt reveals fraud. However, replication attempts themselves rarely, if ever, detect fraud: fraud is usually found through forensic examination of original data (Simonsohn, 2013).

[5] Journals in such a situation may wish to avail themselves of methods that have recently been used to successfully predict replication, such as prediction markets or expert surveys (Dreber et al., 2015; DellaVigna and Pope, 2018; DellaVigna et al., 2019; Landy et al., 2020).

delay publication of the original study. Luckily, the authors of the original study (and co-authors of this study) were willing to push forward with publishing the results, even though their original experiment did not replicate.

3.2 Who Should Pay?

Perhaps the most difficult question for journal-based replication is where money for the replication attempt should come from. The benefit to authors, journals, and the research community does not change based on who pays for a replication attempt. Thus, the preferences of authors and journals about who should pay are opposed. We believe that resolving this opposition likely rests on journals implementing other policies to increase author welfare.

Having the journal pay for replication attempts would likely be quite difficult in practice, although it does honor the principle of increasing author welfare as much as possible. In particular, journals rarely, if ever, apply for research grants. Journals could raise their submission fees, but this would likely lead to a decline in submissions. If submission fees were raised across the board, non-experimental scholars may resent funding public goods for experimental scholars, and may choose to take their work elsewhere. If submission fees were raised only on experimental submissions, this would result in a larger fee raise, and drastically reduce the number of experimental submissions.

On the other hand, asking authors of the original study to pay for replications may also decrease submissions. Such a funding model might look a lot like open-access publishing, where publication fees are charged only for articles that make it through the reviewing process. In the long term, authors could account for this cost in grant applications. As most grant costs cover fixed expenses of experiments—programming, salaries, etc—the marginal increase in funding would likely not be large. However, in the short term this would not be possible. Even in the long term, experimental scholars may try to save money by first submitting to journals that do not require replication fees.

Our suggestion is for journals to commit to not asking for additional experimental treatments as part of the revision process in exchange for authors covering the costs of replication. Currently many editors request additional experimental treatments, and sometimes even reject papers on the outcomes of these additional treatments. As such, the overall effect of these two policies would likely be at most a small cost increase to experimental scholars, and probably a reduction in frustration and time spent on revisions due to requests to add more

experimental treatments.

In the case of our pilot, one of the co-editors of the *Journal of Public Economics* (Snowberg) volunteered his own research funding for the study. Without this commitment, the pilot would not have been possible: the editors of the *Journal of Public Economics* would not have considered allowing the pilot to be run, and frankly, neither would the authors of the original study (Drazen and Ozbay).

3.3 How Exact Should the Replication Attempt Be?

The replication literature generally tries to ensure replication attempts are as close to exact as possible (Open Science Collaboration, 2015).[6] This rule-of-thumb is born of the experience that authors of the original study will often claim their result failed to replicate due to small differences, or changes in "hidden moderators" between the original experiment and the replication attempt (Bargh, 2019). Of course, no replication is truly exact: the replication attempt is never run at the same time as the original experiment, is (almost) never run using the same participants, and is rarely run in the same participant pool. Because of these (usually necessary) differences, efforts are made to make the rest of the replication attempt as similar as possible to the original experiment.

In our case, we struggled with whether or not to replicate one particular element of the original experimental design. The original study assigned treatment at the experimental session level, rather than at the individual level. Changing assignment to the individual level may cause a replication to fail because instructions can no longer be read aloud, as different participants will need different instructions (Chen et al., 2020). This can cause a failure of common knowledge, which can lead to other equilibria being played. On the other hand, the practice of assigning treatment at the session level raises concerns that results may be due to the time of day, or day of week, when an experimental session is conducted, or other factors, rather than due to the treatment itself (Green and Tusicisny, 2012; Ericson, 2018).[7] As far as we are aware, this concern is largely theoretical thus far: while it could be responsible for experimental results, we could find no case where it has been shown that this

[6] In the literature, this is sometimes referred to as a *direct replication*, in order to contrast that approach with a *conceptual replication*. The latter generally refers to running a different experiment that is designed to test the same theory in a similar way. For examples see Landy et al. (2020).

[7] Note that many of these concerns relate to which participants will show up on different days/times, rather than a direct effect of the time the experiment is run.

confound has played a role. However, to account for this concern, we present the results of our replication with multiple approaches to hypothesis testing, two of which are intended to correct for session-level treatments. In addition, while the original study was performed in the U.S., the replication was performed in Spain. This type of discrepancy may be something to avoid in future versions of journal-based replications, as it is possible that our opposite result is due to differences in participant pools.

3.4 Size and Scope of the Replication Attempt

An element of experiments that is usually changed between the replication attempt and the original is the number of participants. As even studies that do replicate often have lower effect sizes in the replication attempt than in the original, this implies that replications with the same number of participants will tend to be underpowered. While there is not a standard for the power of replications—and thus the number of participants—some have suggested that replications should be 2.5 times the original sample size (Simonsohn, 2015), while others have suggested 90% power of detecting an effect size half as large as the original (Camerer et al., 2018). Both approaches will usually lead to replications with sample sizes that are considerably larger than those of original studies.

Cost aside, both journals and authors would prefer a larger replication attempt as both benefit from a successful replication, although authors will usually benefit more. As such, this may be an additional argument for asking authors to pay for replication attempts, as they will then be able to make the tradeoff between budget and size of a replication attempt. We believe this is a secondary consideration in deciding who should pay.

In our case, using a rule-of-thumb of 90% power to defect 50% of the original effect size implied that we could have saved money by running a smaller experiment than the original. However, given concerns we had about session-level treatment, reducing the number of sessions—and thus power in hypotheses tests that correct for the possibility of session-level effects—seemed unwise. Further, our budget constraint did not bind, so we chose instead to keep the number of sessions and participants the same. The original study cost approximately $2,500 in participant compensation, while for the replication the amount was 2,365€ ≈ $2,800. The total cost of the replication, including overhead and administration, was 6,335€ ≈ $7,400.[8] The fact that university labs often do not charge their own researchers

[8]The University of Maryland lab does not directly charge its own experimenters for overhead. The

for overhead and administration is another argument for having authors arrange, and pay for, their own replication attempts, a topic we return to in Section 5.1.

3.5 Which Studies Should be Replicated?

There is an easy, but perhaps impractical, answer to the question of "Which studies should be replicated?" All of them. If a journal believes that a study will have sufficient impact to be published, it seems that the journal should also wish to ensure the result can be replicated.[9]

If, for whatever reason, the financial means to replicate all studies are not available, then the best course of action is likely to choose articles for replication attempts randomly. If authors are paying for replication, this has the benefit of being procedurally fair. If the journal is paying for replication, this will prevent authors from trying to "game" policies dictating which studies are replicated. For example, in our pilot, the editors of the *Journal of Public Economics* did not think we should consider a study with any junior authors. While this was a sound policy for the pilot, as a general matter it would create perverse incentives for authors to represent themselves as having an upcoming career-defining moment in order to avoid having their study replicated. Another related approach would be to run prediction markets on which published studies are likely to replicate, and use the prices for specific studies to change the odds a study is chosen for replication. An example of such a policy would be to give all studies a baseline positive probability of being picked for replication, with those studies for which the information value of a replication is the highest being given a higher probability.[10] Another possibility would be to give papers that have already been

discrepancy between participant compensation in the original and replication is largely due to the fact that we directly translated U.S. \$1 = 1€, although at the time of the experiment U.S. \$1 ≈ 0.85€. The original payment records from the Maryland lab were destroyed for privacy reasons, so our estimate of total participant compensation comes from simulating a random round being paid for each session, and then averaging across simulations. Applying this same method to the Valencia data, we obtained an estimate of 2,337€ versus the 2,365€ we actually paid.

[9] A related question, which is beyond the scope of this work, is how many replication attempts should be made for a single study. Due to financial constraints (both our own, and what we expect would be those of future journal-based replicators) we did not consider more than one replication attempt per experiment. This is a general blind spot ino the literature: Most prior replication attempts also only include a single replication of each experiment. The main exception are the "Many Labs" projects (Klein et al., 2018; Ebersole et al., 2016), in which several labs all replicate the same psychological study. Whether one failed or successful replication attempt is sufficient depends on the trade-off between learning more about a specific study, and learning more about a study without a single replication attempt (Frankel and Kasy, 2018). Some might consider *conceptual replications*—in which some elements of the design are purposely varied—than another direct replication to be more useful in understanding generalizability of results (see, for example, Yarkoni, 2019; Landy et al., 2020).

[10] That is, studies with prices that are close to 50 out of 100, thus an implied probability of replicating of

heavily cited a higher probability of being selected for replication.

Expensive field studies present a gray area. Our argument for journal-based replication, which draws on that of Coffman and Niederle (2015), relies on replications being relatively inexpensive. When they are not, then pre-analysis plans are likely a more cost-effective way of ensuring experimental reliability. Journals could choose not to mandate replications for experiments with a cost over some amount, although this may also create incentives for authors to "game" the policy. An improved policy would be that, for experiments over a certain cost, one of either a pre-analysis plan or a plan for a journal-based replication would be necessary for acceptance. Journal-based replication would be necessary for studies that cost less than this amount, as replication attempts are better at detecting false positives (Coffman and Niederle, 2015).

3.6 What if an Error is Discovered?

It is not unusual for errors to be discovered after a paper is published. This is unsurprising as reviewers rarely delve deeply into the analysis code and data collection underlying studies. Even when these errors can be argued to be insubstantial, their discovery can be embarrassing, and lead to some amount of defensiveness (see, for example, Donohue and Levitt, 2001; Foote and Goetz, 2008; Donohue and Levitt, 2008; Miguel and Kremer, 2004; Aiken et al., 2015; Hicks et al., 2015). However, when these errors are more substantive, this can lead to cycles of recriminations, accusations of bad faith, and may not ultimately change people's posteriors as much as if the error had never been introduced into the literature.[11]

Such errors may not be as common in experimental work, as analyses tend to be less complicated, and there are rarely settled "correct" ways to implement particular elicitations or treatments (see, for example, Andreoni et al., 2015; Augenblick et al., 2015). Yet, errors may still occur and roil a literature for years (see, for example, Rand et al., 2012; Bouwmeester et al., 2017). Wherever one comes out on these debates, all involved would have probably preferred that agreed upon errors be caught before the original paper was published, and that questions about certain results had been appropriately discussed.

50%. It is worth noting that so far prediction markets have mainly been run on published studies, and it is unclear how well they would work on selecting studies that are being considered for publication (Dreber et al., 2015, see, for example).

[11] See, for example, (Hoxby, 2000; Rothstein, 2007; Hoxby, 2005; Acemoglu et al., 2001; Albouy, 2012; Acemoglu et al., 2012).

Errors in analysis found in a journal-based replication should be dealt with as quickly as possible, although, in general, we do not think they should impact publication. If the error is substantively meaningful—for example, changing a statistically significant effect to a null effect—this may warrant a complete re-writing around the result, and suspending the replication attempt. Under the principle of ensuring that authors receive as much surplus as practical, we would advocate for publishing the null result. However, a note should be made that the original paper contained a substantial error in analysis, as these instances should be known, and possibly impact authors' reputations. Errors that are not substantively meaningful should be corrected in the final version of the paper (which, as noted above, we believe should include both the original and replication experiments). Most editors have seen many such small errors corrected in the review process, and they are not normally commented upon in the final paper (except, occasionally, for some small word of thanks from the authors).

Errors in implementation are more difficult to address, in part because experimental practice is usually the topic of lively debate, so there is rarely a "correct" implementation. As such, in most cases, it will be difficult to determine if a particular dispute is due to an error, or simply a difference of opinion. An exception is if the description of an experimental technique in a manuscript does not match what was actually done in the experiment. Is this case, the paper should be corrected, and if the mismatch is substantial enough, the paper should not proceed to replication without an additional round of review (with the same reviewers, or possibly new ones). The editor may even judge that s/he would have not sent the paper out for review had the experiment been properly represented, and reject it.

4 Replication Results

We now present an overview of the experiment and our replication results, before discussing how the experience of conducting a pilot of journal-based replication might inform a sustainable journal-based-replication policy.

4.1 Experimental Overview

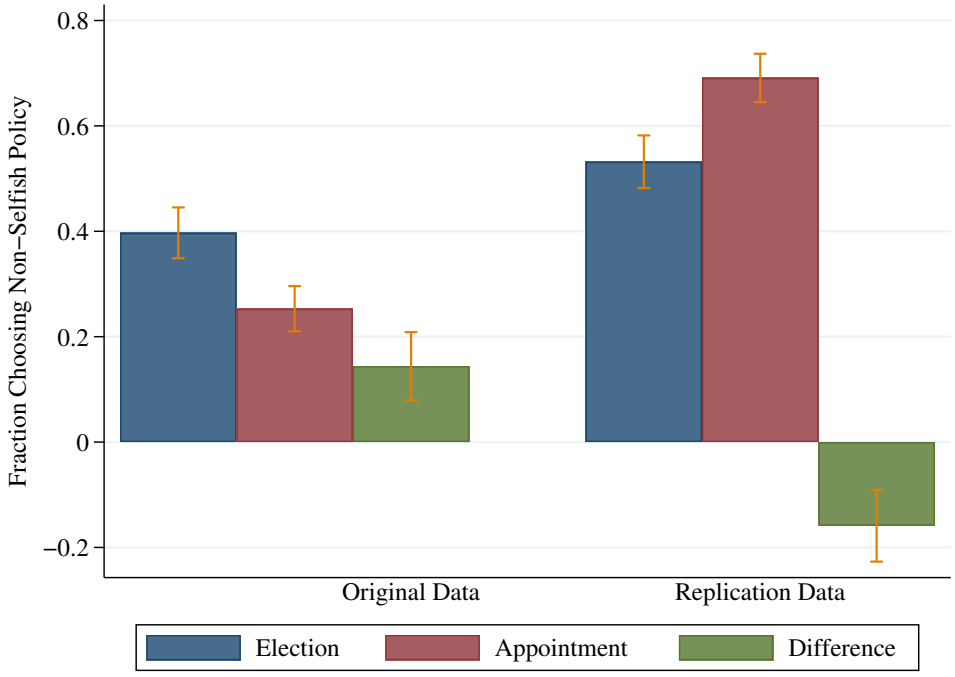
The main goal of the experiment in Drazen and Ozbay (2019) was to test the idea that, due to reciprocity, elected leaders might be more willing to act *non-selfishly*—that is, implement

a policy further from their own ideal policy—than appointed leaders. The experimental design is succinctly described in the original paper, which we quote here:

> At the beginning of each session, each subject was randomly assigned one of two roles: "candidate" or "citizen." There were twice as many candidates as citizens. The assigned roles stayed fixed for all 20 rounds (until the end of the experiment). At the beginning of each of the 20 rounds in a session, all participants were randomly put into groups of 3 people. Hence, there could be no "reputation" effects as the session proceeded. Each group consisted of two candidates and one citizen. Independent from the assigned role (candidate and citizen), every participant was randomly assigned a *type* in each round. A type was any integer number from 0 to 100 drawn from a uniform distribution, which is essentially the participant's most preferred policy. Unlike the fixed roles, assigned types changed from one round to the next. We balanced the random draws by using the same sequence of random numbers for each treatment, so the random value draws for each session in the Election Treatment are matched with the random draws for the corresponding session of the Appointment Treatment.
>
> After being informed about the type of each candidate, in the Election Treatment, the citizen choses one of the candidates. In the Appointment Treatment, one of the candidates was randomly appointed. The elected candidate in the Election Treatment, or the appointed candidate in the Appointment Treatment, was informed about the types of both the opponent candidate and the citizen and was then given the authority to decide which policy would be implemented. A policy was required to be an integer number from 0 and 100, where individuals learned the outcome of each round before the next took place.
>
> Earnings in each round depended on the distance between type and policy. Formally, the earnings in a round were $100 - |\text{TYPE} - \text{POLICY}|$ Experimental Currency Units (ECU) where 1 USD = 5 ECU. It is important to note that all participants, both citizens and candidates, have their earnings computed in this fashion, and the policy choice of the winning candidate affected the earnings of both opponent candidate and the citizen. Once all 20 rounds were finished, one round out of the 20 was randomly picked, and the earnings in that round were the final earnings of the experiment in addition to a $5 participation fee.

The main outcomes analyzed both in the original paper and here are the fraction of leaders acting non-selfishly, and the *magnitude* of the deviation from selfish behavior—that is, the absolute value of the difference between the policy chosen and the leader's ideal policy. As in our replication, the experiment in Drazen and Ozbay (2019) had 120 participants across

Figure 1: Our replication obtained the opposite result from the original study.



Notes: Fraction of leaders choosing a policy other than their most preferred option, with 95% confidence intervals.

8 sessions (4 election and 4 appointment), although those subjects were recruited at the University of Maryland, as opposed to our study, which took place at the University of Valencia (Spain). Both the study and the replication were conducted in English.

4.2 Main Finding

The main finding of Drazen and Ozbay (2019) is that those in the Election Treatment were significantly more likely to act non-selfishly than those in the Appointment Treatment. The main result in our replication was the opposite: namely, those in the Appointment Treatment were significantly more likely to act non-selfishly, as shown in Figure 1. Additionally, the overall rates of non-selfish behavior are higher in the replication data in both treatments.

There is some disagreement in the literature about how to conduct hypothesis testing in data from experiments with session-level treatment. In particular, if one believes that

the time of day or day of week of the experimental session (through either time effects, population effects, etc.) does not systematically affect experimental treatments, then the classical approach of assuming spherical errors is appropriate (Cameron et al., 2008; Young, 2019). If, on the other hand, one is concerned about such effects, then some approach should be taken to recognize the effect of intra-session correlation in the error terms (Green and Tusicisny, 2012; Abadie et al., 2017; Ericson, 2018). The typical approach is to cluster standard errors at the session level. However, as clustered standard errors have heavy tails, especially with small numbers of clusters, this may leads to more false positives than classical standard errors (Young, 2019). As such, we implement two approaches that are gaining popularity in the literature: randomization inference with clustering (Young, 2019), and Wild-Bootstrapping (MacKinnon and Webb, 2018). Both of these approaches appear to be less biased when there are small numbers of clusters.

We present the results underlying Figure 1 in Table 1, with different p-values corresponding to the three different forms of hypothesis testing described above. In all cases, the treatment effect of switching from election to appointment is highly statistically significant when using classical standard errors. Using either randomization inference or wild bootstrapping tends to maintain statistical significance, but at a level of around 0.05 or 0.1. P-values are likely higher in the replication sample due to higher intra-cluster correlation in that data. Regardless of the method used, the difference between the two treatment effects is statistically significant at conventional levels.

There are two plausible interpretations for why our replication obtained the opposite result of the original study. The first plausible explanation is that there is a difference in culture between the U.S. and Spain (or, more precisely, between the two participant populations drawn from those two countries) leading to opposite results. There is some tentative support for this in the literature, as other studies have also found a difference in reciprocity between participants in Spain and those in other OECD countries (Georgantzis et al., 2013; Waichman et al., 2015). Based on this literature, this difference in reciprocity across coun-

Table 1: Do leaders behave non-selfishly? It depends on the population.

	Fraction of Leaders	Magnitude of Effect
Panel A: Original Data		
Election	0.40	4.67
$N = 398$	(0.025)	(0.59)
Appointment	0.25	2.25
$N = 395$	(0.022)	(0.39)
Treatment Effect	0.14	2.42
	(0.033)	(0.71)
p-values		
Classical	$p = 0.000$	$p = 0.001$
Wild Bootstrap	$p = 0.043$	$p = 0.014$
Randomization Inference	$p = 0.060$	$p = 0.058$
Panel B: Replication Data		
Election	0.53	12.3
$N = 385$	(0.026)	(1.04)
Appointment	0.69	20.2
$N = 392$	(0.023)	(1.28)
Treatment Effect	-0.16	-7.89
	(0.035)	(1.65)
p-values		
Classical	$p = 0.000$	$p = 0.000$
Wild Bootstrap	$p = 0.087$	$p = 0.10$
Randomization Inference	$p = 0.087$	$p = 0.11$
Panel C: Difference between Treatment Effects		
Difference	0.30	10.3
	(0.048)	(1.78)
p-values		
Classical	$p = 0.000$	$p = 0.000$
Wild Bootstrap	$p = 0.005$	$p = 0.037$
Randomization Inference	$p = 0.003$	$p = 0.002$

Notes: Standard errors in parentheses. Wild Bootstrapped and Randomization Inference p-values from 10,000 draws.

tries may explain the difference in results. One could also posit other cultural factors may be responsible, although we have not found specific evidence to support this interpretation. For example, it may be that appointed (versus elected) leaders view their responsibilities towards their constituent populations differently in Spain than in other OECD countries.[12] The second explanation is that both findings are statistical noise, and that election and appointment do not have any systematic effect in either country on pro-social behavior in this particular experiment.

The remainder of the results from the original study can be found in Online Appendix A. We present all three forms of hypothesis tests, similar to Table 1. Both datasets exhibit a decrease in non-selfish behavior in later rounds (Table 2). However, in our replication data, the difference between the two treatments is almost entirely driven by differences during the last ten rounds (rounds 11–20, as shown in Table 3). To put this another way, in the original data, the decrease in non-selfish behavior is relatively even across both treatments. Instead, in the replication data, the decrease in non-selfish behavior is far more pronounced for those in the Election Treatment. The remainder of the specifications in Drazen and Ozbay (2019) are dedicated to testing how particular theories of reciprocity or other-regarding preferences might explain the main result (Tables 4–8). Given that the main result is different in the replication dataset, these tests are no longer well specified. However, for completeness, we show them as well. As the results in the original paper only contained hypothesis tests using classical standard errors, we also reproduce the tables from the original paper with all three types of hypothesis tests described above in Online Appendix B.

5 What we Learned

[12]Differences may also be attributable directly to the particular participants and conditions in the original study or replication: for example, the participants' level of experience with lab experiments, how well they understood the instructions, how approachable the facilitators were, etc. These sorts of "hidden moderator"-type explanations for differences between original and replication studies have found little support in the literature (Inbar, 2016; Snowberg and Yariv, 2021).

The basic lessons of our pilot were that journal-based replication is relatively easy to implement and execute, but that care must be taken to ensure its speedy conclusion. This second basic lesson—as well as other factors—caused us to think more deeply about specific elements of journal-based replication, which we discuss in this section.

5.1 How Exact Should the Replication Attempt Be?

Getting a result that is statistically significant in the opposite direction of the original result is rare in replication attempts. This potentially indicates that despite our attempt to make our replication as "exact" as possible–see Section 3.3—using a different participant pool may have been a more important change than we recognized. This fact, plus a comment from a helpful reviewer, has lead us to believe that authors conducting their own replications may be a viable option. Here we evaluate that possibility.

Timing of the replication attempt is likely more important than who actually implements it. Much of the replication literature (and our own study) is focused on the problems stemming from "researcher degrees of freedom" or specification searches and post-hoc rationalization, including "p-hacking" and "forking" and the fact that "p-hacked" and "forked" studies are less likely to be robust (Leamer, 1983; Simmons et al., 2011; Gelman and Loken, 2013). If one runs two instantiations of the same experiment, one could simply "p-hack" the results of both studies to find a statistically significant effect that exists in both studies. If the replication is run after the original paper is vetted, then this is not a concern. In particular, the fact that the paper would already be accepted before the replication attempt is made would reduce (but probably not eliminate) the moral hazard to try to make the original experiment replicate.

There may be benefits in allowing the original authors to replicate their own study. The primary benefit is that it would allow access to the same participant pool and environment, thus reducing potential confounds. On the other hand, some scholars view replication as ensuring results are not particular to a participant pool that is "special" in some way. The

importance of this consideration depends on the intended purpose of replication, as well as in the amount of heterogeneity one believes there is between labs and in other experiment-level variables—an active topic of discussion in the literature (Stanley et al., 2018; Kenny and Judd, 2019; McShane et al., 2019; Simmons and Simonsohn, 2019). Allowing authors to replicate their own studies would also likely be cheaper, as it would allow authors to avoid the overhead charges from contract laboratories.[13] Additionally, any publication delays due to slow execution of a replication or write up would be largely the responsibility of the authors, sparing the journal from some administrative headaches. Finally, should the results fail to replicate, this would avoid conflict between the original team and the replication team.

Weighing against allowing authors conducting their own replications would be the lack of an "additional set of eyes" that might discover inadvertent errors, or even outright fraud. It is worth noting that this latter concern is unlikely to be substantially affected by allowing authors to do their own replications. As the paper has already been accepted, there would be less motivation for fraud than with an initial submission. Moreover, replication attempts themselves rarely, if ever, detect fraud: fraud is usually found through forensic examination of original data (Simonsohn, 2013).

5.2 What Should be Done with the Results?

Our experience with journal-based replication leads us to strongly believe that the results of the replication should be included in the same paper as the original experiment. Without this commitment, other scholars may never learn if a given study replicates or not.

As noted above, conditionally accepting a manuscript before replication is attempted does not imply that the paper needs to be published "as is." The experience of running a pilot allows us to make more concrete recommendations about how the replication should affect the published paper. In our opinion, if the results of the replication attempt confirm the original experimental results, then this can be noted, and the data can be pooled in the

[13] Another option would be that the authors replicate their study in an independent lab, as in, for example, Gaechter et al. (2019).

analysis. If the replication attempt finds null results, this should result in a thorough re-writing of the paper, focusing on the replication, with the original results effectively serving as a pre-analysis plan that can either be discussed, or relegated to an appendix. If the replication produces the opposite result, then both sets of results should be reported, with the authors encouraged to try to explain the conflicting results. To a limited extent, we have tried to do this in the current paper.

5.3 What if an Error is Discovered?

During the replication process we encountered a clear, but insubstantial error, and one that were less clear but more important. This lead us to think more about the importance of an "additional set of eyes" in replications, and about clear policies for dealing with errors.

The clear, but ultimately insubstantial, error was that the original paper had used a mix of standard OLS regressions for coefficients, and random effects GLS analysis for standard errors. This was corrected in the revision process of the original paper, with the results being fairly similar.

More important, but less clearly erroneous, features were addressed in revising this paper. With the help of the editor, we came to believe it was important to present hypothesis tests that allowed for the possibility of session-level effects. This is discussed extensively in Section 4.2. Additionally, we changed specifications involving the fraction of participants acting non-selfishly from Logit to a linear-probability OLS model. This is based on the fact that measurement error in left-hand-side variables can lead to biases in estimation results when using discrete-choice models, but not when using linear-probability models in OLS (Gillen et al., 2019). While these features might be described as older research practices, as opposed to errors, in the spirit of disclosure described in Section 3.6, we have documented and described them here.

While these changes emphasize the value of an "additional set of eyes" on papers, it is not clear to us that most journal-based replications would benefit as much as we have. First,

writing this paper subjected the results of the original paper to the scrutiny of an additional set of authors, and perhaps more importantly, research assistants. Moreover, the fact that this paper was reviewed separately from the original gave us access to the suggestions of an additional set of reviewers, and an editor to corral and amend those suggestions. These additional resources would not generally be part of a standard journal-based replication.

Thus, we believe the most important issue highlighted by our experience was the need for clear processes and procedures before undertaking a journal-based replication. We lacked many of these (includingg what to do if errors were discovered), and this lead to delays and extensive discussions as we wrestled with what to do in various circumstances. This was, of course, by design, as the entire point of the proof-of-concept was to attempt to learn what issues we may have overlooked. However, it is still worth emphasizing the importance of setting policies—and the appropriate level of editorial discretion—to address as many potential issues as possible ahead of time.

5.4 Who Should Pay?

This brings us back to what we earlier labeled the most difficult question about journal-based replication, "Who should pay?" While our pilot provided little direct information about this question, some of the information in this section did influence our thinking indirectly.

Having authors pay for the replication is a natural complement to allowing them to undertake the replication attempt themselves, as discussed in Section 5.1. As noted there, this would put responsibilities for delays largely in the hands of the authors. Moreover, it would also give authors discretion over another potentially contentious aspect of the replication attempt—the sample size, and thus power, as discussed in Section 3.4. However, allowing authors to conduct their own replication attempt may reduce the chance that the replication will find or document any potential flaws, as described in Section 5.3.

As noted earlier, having authors pay for replication attempts would likely require the journal to make some concessions in order to avoid a reduction in submissions. We believe

that promising that additional treatments would not be required as part of the normal review process, as described in Section 3.2, may be sufficient. A hybrid may also be useful: upon submission, authors could decide if they wished to be part of the journal-based-replication track, in which case no additional treatments would be requested; or, part of a more standard track where additional treatments may be required (and would not necessarily be sufficient for publication).[14] However, there are likely other creative ways to partially compensate authors for bearing the monetary cost of replication.

6 Discussion

This paper proposes and executes a pilot of a novel mechanism of ensuring replication: journal-based replication. By making publication decisions before a replication is conducted, this mechanism would reduce the possibility of "file-drawer" problems and publication bias. The main alternative, at this point, is registered reports, described in Section 2.1 in which a paper is accepted on the basis of an experimental design and a pre-analysis plan (Nosek and Lakens, 2014). Compared to this alternative, journal-based replication allows editors and reviewers the ability to see the experimental results from one set of participants before making a publication decision. As a finding's usefulness, and citation count, often depends on what, exactly, that finding is, this allows editors and reviewers to make a more informed judgement about the potential impact of a paper (Coffman and Niederle, 2015; Brodeur et al., 2016). However, a hybrid between journal-based replication and registered reports that includes both experimental data and an experimental design and analysis plan to replicate and extend those results, may be an improvement on both. For example, as part of the initial submission of an experimental paper, authors could specify the details of the replication they would run—the sample, size, etc.—if the paper were to be accepted. If editors looked

[14]Colin Camerer suggested this to us based on his own (unfortunately unsuccessful) grant application to conduct a wide-ranging journal-based replication campaign. He argued that, in equilibria, the choice of track would also serve as a signal of the original experiment's quality. While we believe it is more likely to be a signal of the author's financial resources, it is an idea worth exploring, as it may also force authors to focus their resources on fewer, better-designed experiments.

favorably upon submissions with such a commitment to replication, this could create a virtuous cycle.

The biggest concern for implementing a journal-based replication policy is the source(s) of funding. Aside from this concern, both the journal and the authors found the process to be straightforward, with the writing up of these results to be far more time-consuming than conducting the replication itself.[15] Comparing the model of journal-based replication to a model in which replication occurs through other means, the difference would be that the costs of replication would be borne by the journal, and, more likely, the authors of the original study. However, as the benefit of studies primarily accrues to the authors and the journals that publish them, we believe these parties should bear the costs of replication. Moreover, if replication were seen as a standard part of the research process, we hope that granting agencies would adjust their funding accordingly, although we suspect this might result in fewer grants rather than an increase in the pie, at least in the short to medium term. Finally, as people wonder what the purpose of journals is in an age of open access (Resnick and Belluz, 2019), it seems that enforcing replication could be one such purpose.

Our proof-of-concept provided information on a number of practical problems that may arise in actually trying to implement a journal-based replication policy. The fact that our replication results were the *opposite* of the original results lead us to struggle with a number of unforeseen questions, and illustrates the usefulness of replication. In particular, whether one believes that our opposite finding is due to broad cultural differences, or statistical noise, likely depends on one's prior. If one has seen a number of failed replications, it is likely that person would presume this to be just another, somewhat atypical, failure. Or, one may see a broader pattern in the literature, and believe that understanding how this experiment, and others that rely on reciprocity, vary across cultures seems like a fruitful topic for further research. As reciprocity has been found to be important in, for example, corruption and voting behavior (for example, see, Finan and Schechter, 2012), there may also be implications

[15]Funding was also never a question, due to Snowberg's Canada Excellence Research Chair grant, but this would not typically be the case.

for how institutions function across these different cultures and countries. Regardless of one's posterior on this question, both are informed by the result of the replication attempt. This is precisely the point of replications—they add data to inform the probability that a particular hypothesis is correct.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge, "When should you Adjust Standard Errors for Clustering?," 2017. National Bureau of Economic Research Working Paper #24,003.

Acemoglu, Daron, Simon Johnson, and James A. Robinson, "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 2001, *91* (5), 1369–1401.

_ , _ , and _ , "The Colonial Origins of Comparative Development: An Empirical Investigation: Reply," *American Economic Review*, 2012, *102* (6), 3077–3110.

Aiken, Alexander M., Calum Davey, James R. Hargreaves, and Richard J. Hayes, "Re-analysis of Health and Educational Impacts of a School-based Deworming Programme in Western Kenya: A Pure Replication," *International Journal of Epidemiology*, 2015, *44* (5), 1572–1580.

Albouy, David Y., "The Colonial Origins of Comparative Development: An Empirical Investigation: Comment," *American Economic Review*, 2012, *102* (6), 3059–3076.

Andreoni, James, Michael A. Kuhn, and Charles Sprenger, "Measuring Time Preferences: A Comparison of Experimental Methods," *Journal of Economic Behavior & Organization*, 2015, *116*, 451–464.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger, "Working over time: Dynamic inconsistency in real effort tasks," *The Quarterly Journal of Economics*, 2015, *130* (3), 1067–1115.

Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg, "Decision Theoretic Approaches to Experiment Design and External Validity," in Esther Duflo and Abhijit Banerjee, eds., *Handbook of Field Experiments*, Elsevier, 2017, pp. 141–174.

_ , _ , Sergio Montero, and Erik Snowberg, "A Theory of Experimenters: Robustness, Randomization, and Balance," *American Economic Review*, 2020, *110* (4), 1206–1230.

Bargh, John, "Comment on 'Now John Bargh's Famous Hot-Coffee Study Has Failed To Replicate'," 2019. https://digest.bps.org.uk/2019/01/02/now-john-barghs-famous-hot-coffee-study-has-failed-to-replicate/#comment-63119.

Berry, James, Lucas C. Coffman, Douglas Hanley, Rania Gihleb, and Alistair J. Wilson, "Assessing the Rate of Replication in Economics," *American Economic Review*, 2017, *107* (5), 27–31.

Bouwmeester, Samantha, Peter P.J.L. Verkoeijen, Balazs Aczel, Fernando Barbosa, Laurent Bègue, Pablo Brañas-Garza, Thorsten G.H. Chmura, Gert Cornelissen, Felix S. Døssing, Antonio M. Espín et al., "Registered Replication Report: Rand, Greene, and Nowak (2012)," *Perspectives on Psychological Science*, 2017, *12* (3), 527–542.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg, "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, 2016, *8* (1), 1–32.

Butera, Luigi and John A. List, "An Economic Approach to Alleviate the Crises of Confidence in Science: With an Application to the Public Goods Game," Technical Report, National Bureau of Economic Research Working Paper #23335 2017.

_ , Philip J. Grossman, Daniel Houser, John A List, and Marie-Claire Villeval, "A New Mechanism to Alleviate the Crises of Confidence in Science—With An Application to the Public Goods Game," Technical Report, National Bureau of Economic Research Working Paper #26801 2020.

Camerer, Colin F. Anna Dreber, and Magnus Johannesson, "Replication and Other Practices for Improving scientific Quality in Experimental Economics," in Arthur Schram and Aljaž Ule, eds., *Handbook of Research Methods and Applications in Experimental Economics*, Edward Elgar Publishing, 2019, pp. 83–103.

_ , _ , Eskil Forsell, Teck-Hua Ho et al., "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 2016, *351* (6280), 1433–1436.

_ , _ , Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer et al., "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour*, 2018, *2* (9), 637.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

Chen, Roy, Yan Chen, and Yohanes E. Riyanto, "Best Practices in Replication: A Case Study of Common Information in Coordination Games," *Experimental Economics*, 2020, pp. 1–29.

Christensen, Garret and Edward Miguel, "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 2018, *56* (3), 920–980.

Coffman, Lucas C. and Muriel Niederle, "Pre-analysis Plans have Limited Upside, Especially where Replications are Feasible," *Journal of Economic Perspectives*, 2015, *29* (3), 81–98.

Deaton, Angus, "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, June 2010, *48* (2), 424–455.

DellaVigna, Stefano and Devin Pope, "What Motivates Effort? Evidence and Expert Forecasts," *The Review of Economic Studies*, 2018, *85* (2), 1029–1069.

_ , _ , **and Eva Vivalt**, "Predict Science to Improve Science," *Science*, 2019, *366* (6464), 428–429.

Donohue, John J. and Steven D. Levitt, "The Impact of Legalized Abortion on Crime," *The Quarterly Journal of Economics*, 2001, *116* (2), 379–420.

_ **and** _ , "Measurement Error, Legalized Abortion, and the Decline in Crime: A Response to Foote and Goetz," *The Quarterly Journal of Economics*, 2008, *123* (1), 425–440.

Drazen, Allan and Erkut Y. Ozbay, "Does 'Being Chosen to Lead' Induce Non-selfish Behavior? Experimental Evidence on Reciprocity," *Journal of Public Economics*, 2019, *174* (1), 13–21.

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson, "Using Prediction Markets to Estimate the Reproducibility of Scientific Research," *Proceedings of the National Academy of Sciences*, 2015, *112* (50), 15343–15347.

Ebersole, Charles R. Olivia E. Atherton, Aimee L. Belanger, Hayley M. Skulborstad et al., "Many Labs 3: Evaluating Participant Pool Quality across the Academic Semester via Replication," *Journal of Experimental Social Psychology*, 2016, *67*, 68–82.

Ericson, Keith M., "Design Issues in Economics Lab Experiments: Randomization," 2018. https://practicingeconomist.com/2018/03/07/design-issues-in-economics-lab-experiments-randomization/.

Finan, Frederico and Laura Schechter, "Vote-buying and Reciprocity," *Econometrica*, 2012, *80* (2), 863–881.

Foote, Christopher L. and Christopher F. Goetz, "The Impact of Legalized Abortion on Crime: Comment," *The Quarterly Journal of Economics*, 2008, *123* (1), 407–423.

Frankel, Alexander and Maximilian Kasy, "Which Findings should be Published?," 2018. Harvard University, *mimeo.*

Gaechter, Simon, Chris Starmer, and Fabio Tufano, "The Surprising Capacity of the Company You Keep: Revealing Group Cohesion as a Powerful Factor of Team Production," 2019. CeDEx Discussion Paper Series No. 2019-16.

Gelman, Andrew and Eric Loken, "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There is no 'Fishing Expedition' or 'p-Hacking' and the Research Hypothesis was Posited Ahead of Time," 2013. Columbia University, *mimeo.*

Georgantzis, Nikolaos, Juan A. Lacomba, Francisco Lagos, and Juliette Milgram, "Trust and Reciprocity among Mediterranean Countries," 2013. Universitat Jaume I, *mimeo.*

Gillen, Ben, Erik Snowberg, and Leeat Yariv, "Experimenting with Measurement Error: Techniques and Applications from the Caltech Cohort Study," *Journal of Political Economy*, 2019, *127* (4), 1826–1863.

Gneezy, Uri and Marta Serra-Garcia, "Non-replicable Publications Are Cited More than Replicable Ones," 2020. University of California, San Diego, *mimeo.*

Green, Donald P. and Andrej Tusicisny, "Statistical Analysis of Results from Laboratory Studies in Experimental Economics: A Critique of Current Practice," 2012. Columbia University, *mimeo.*

Hicks, Joan H., Michael Kremer, and Edward Miguel, "Commentary: Deworming Externalities and Schooling Impacts in Kenya: A Comment on Aiken et al. (2015) and Davey et al. (2015)," *International Journal of Epidemiology*, 2015, *44* (5), 1593–1596.

Hoxby, Caroline M., "Does Competition among Public Schools Benefit Students and Taxpayers?," *American Economic Review*, 2000, *90* (5), 1209–1238.

_ , "Competition Among Public Schools: A Reply to Rothstein (2004)," 2005. National Bureau of Economic Research Working Paper #11216.

Imbens, Guido W., "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, June 2010, *48* (2), 399–423.

Inbar, Yoel, "Association Between Contextual Dependence and Replicability in Psychology may be Spurious," *Proceedings of the National Academy of Sciences*, 2016, *113* (34), E4933–E4934.

Kasy, Maximilian, "Why Experimenters Might not Always want to Randomize, and what they Could do Instead," *Political Analysis*, 2016, *24* (3), 324–338.

Kenny, David A and Charles M Judd, "The Unappreciated Teterogeneity of Effect Sizes: Implications for Power, Precision, Planning of Research, and Replication," *Psychological Methods*, 2019, *24* (5), 578–589.

Klein, Richard A. Michelangelo Vianello, Fred Hasselman, Byron G. Adams et al., "Many Labs 2: Investigating Variation in Replicability across Samples and Settings," *Advances in Methods and Practices in Psychological Science*, 2018, *1* (4), 443–490.

Landy, Justin F., Miaolei Liam Jia, Isabel L. Ding, Domenico Viganola, Warren Tierney, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Charles R. Ebersole, Quentin F. Gronau et al., "Crowdsourcing Hypothesis Tests: Making Transparent how Design Choices Shape Research Results," *Psychological Bulletin*, 2020, *146* (5), 451–479.

Leamer, Edward E., "Let's Take the Con Out of Econometrics," *The American Economic Review*, 1983, *73* (1), 31–43.

MacKinnon, James G. and Matthew D. Webb, "The Wild Bootstrap for Few (Treated) Clusters," *The Econometrics Journal*, 2018, *21* (2), 114–135.

Maniadis, Zacharias, Fabio Tufano, and John A. List, "How to make Experimental Economics Research more Reproducible: Lessons from other Disciplines and a New Proposal," in Cary A. Deck, Enrique Fatas, and Tanya Rosenblat, eds., *Replication in Experimental Economics*, Vol. 18 of *Research in Experimental Economics*, Emerald Group Publishing Limited, 2015, pp. 215–230.

_ , _ , and _ , "To Replicate or not to Replicate? Exploring Reproducibility in Economics through the lens of a Model and a Pilot Study," *The Economic Journal*, 2017, *127* (605), F209–F235.

McShane, Blakeley B., Jennifer L. Tackett, Ulf Böckenholt, and Andrew Gelman, "Large-scale Replication Projects in Contemporary Psychological Research," *The American Statistician*, 2019, *73* (sup1), 99–105.

Miguel, Edward and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 2004, *72* (1), 159–217.

Nosek, Brian A and Daniël Lakens, "A Method to Increase the Credibility of Published Results," *Social Psychology*, 2014, *45* (3), 137–141.

Nosek, Brian A. and Timothy M. Errington, "What is Replication?," *PLOS Biology*, 2020, *18* (3), e3000691.

Olson, Mancur, *The Logic of Collective Action: Public Goods and the Theory of Groups*, Harvard University Press, 1965.

Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science*, 2015, *349* (6251), aac4716.

Rand, David G., Joshua D. Greene, and Martin A. Nowak, "Spontaneous Giving and Calculated Greed," *Nature*, 2012, *489* (7416), 427–430.

Resnick, Briand and Julia Belluz, "The War to Free Science," *Vox*, July 10 2019, *https://www.vox.com/the-highlight/2019/6/3/18271538/open-access-elsevier-california-sci-hub-academic-paywalls.*

Rothstein, Jesse, "Does Competition Among Public Schools Benefit Students and Taxpayers? Comment," *American Economic Review*, 2007, *97* (5), 2026–2037.

Schafmeister, Felix, "The Effect of Replications on Citation Patterns: Evidence From a Large-Scale Reproducibility Project," 2020. Stockholm School of Economics, *mimeo.*

Simmons, Joseph P and Uri Simonsohn, "[76] Heterogeneity Is Replicable: Evidence From Maluma, MTurk, and Many Labs," *DataColada*, 2019.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn, "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological science*, 2011, *22* (11), 1359–1366.

Simonsohn, Uri, "Just Post It: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone," *Psychological Science*, 2013, *24* (10), 1875–1888.

_ , "Small Telescopes: Detectability and the Evaluation of Replication Results," *Psychological Science*, 2015, *26* (5), 559–569.

Snowberg, Erik and Leeat Yariv, "Testing the Waters: Behavior across Participant Pools," *American Economic Review*, 2021, *111* (2).

Stanley, Tom D., Evan C. Carter, and Hristos Doucouliagos, "What Meta-analyses Reveal about the Replicability of Psychological Research," *Psychological Bulletin*, 2018, *144* (12), 1325.

Waichman, Israel, Ch'ng Siang, Till Requate, Aric Shafran, Eva Camacho-Cuena, Yoshio Iida, and Shosh Shahrabani, "Reciprocity in Labor Market Relationships: Evidence from an Experiment across High-income OECD Countries," *Games*, 2015, *6* (4), 473–494.

Yang, Yang, Wu Youyou, and Brian Uzzi, "Estimating the Deep Replicability of Scientific Findings using Human and Artificial Intelligence," *Proceedings of the National Academy of Sciences*, 2020, *117* (20), 10762–10768.

Yarkoni, Tal, "The Generalizability Crisis," *University of Texas at Austin, mimeo.*, 2019.

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," *The Quarterly Journal of Economics*, 2019, *134* (2), 557–598.

Online Appendix—Not Intended for Publication

A Additional Replication Results

All tables in this online appendix correspond to tables in the paper by Drazen and Ozbay (2019), with numbers corresponding to those in the original paper. Data in these tables come from our replication attempt conducted in Valencia, Spain. Differences and similarities between the outcomes in these tables, and those in the paper, are discussed in Section 4.2. Additionally, while some specifications in Drazen and Ozbay (2019) used a discrete-choice Logit model, all of these analyses here use OLS.

Table A.2: The impact of being elected on choosing a non-selfish policy ($N = 777$).

Panel A: Fraction Non-Selfish					
Election	−0.16	−0.16	−0.16	−0.15	−0.15
	(0.035)	(0.035)	(0.035)	(0.036)	(0.035)
p-values					
Classical	$p = 0.000$	$p = 0.000$	$p = 0.000$	$p = 0.000$	$p = 0.000$
Wild Bootstrap	$p = 0.088$	$p = 0.083$	$p = 0.083$	$p = 0.10$	$p = 0.095$
Randomization Inference	$p = 0.084$	$p = 0.089$	$p = 0.085$	$p = 0.11$	$p = 0.11$
Leader's type		0.0003	0.0003	0.0003	0.0005
		(0.0006)	(0.0006)	(0.0006)	(0.0006)
Losing candidate's type			0.0003	0.0003	0.0001
			(0.0006)	(0.0006)	(0.0006)
Citizen's type			−0.0008	−0.0008	−0.0002
			(0.0006)	(0.0006)	(0.0006)
Leader being the closest				−0.037	−0.018
				(0.037)	(0.036)
Period					−0.020
					(0.0030)
Constant	0.69	0.68	0.70	0.72	0.89
	(0.024)	(0.039)	(0.058)	(0.061)	(0.065)
Panel B: Magnitude					
Election	−7.9	−8.0	−7.9	−7.0	−7.2
	(1.7)	(1.6)	(1.6)	(1.7)	(1.7)
p-values					
Classical	$p = 0.000$	$p = 0.000$	$p = 0.000$	$p = 0.000$	$p = 0.000$
Wild Bootstrap	$p = 0.11$	$p = 0.10$	$p = 0.11$	$p = 0.12$	$p = 0.12$
Randomization Inference	$p = 0.11$	$p = 0.11$	$p = 0.12$	$p = 0.19$	$p = 0.20$
Leader's type		0.10	0.10	0.10	0.11
		(0.028)	(0.029)	(0.028)	(0.028)
Losing candidate's type			0.058	0.058	0.050
			(0.029)	(0.029)	(0.028)
Citizen's type			−0.042	−0.042	−0.019
			(0.029)	(0.029)	(0.029)
Leader being the closest				−3.6	−2.8
				(1.7)	(1.7)
Period					−0.85
					(0.14)
Constant	20	15	14	16	23
	(1.2)	(1.8)	(2.7)	(2.9)	(3.1)

Notes: Standard errors in parentheses. Leader being the closest is the dummy variable that indicates that the absolute difference between the leader's type and the ordinary citizen's type is less than the absolute difference between the losing candidate's type and the citizen's type.

Table A.3: The impact of early vs late periods on choosing a non-selfish policy.

	Early	Late	Difference
	Panel A: Fraction Non-Selfish		
Election	0.71	0.35	−0.36
	(0.033)	(0.035)	(0.049)
	p-values		
	Classical		$p = 0.000$
	Wild Bootstrap		$p = 0.020$
	Randomization Inference		$p = 0.000$
Appointment	0.74	0.65	−0.090
	(0.032)	(0.034)	(0.044)
	p-values		
	Classical		$p = 0.054$
	Wild Bootstrap		$p = 0.041$
	Randomization Inference		$p = 0.060$
Difference	−0.026	−0.29	
	(0.049)	(0.044)	
p-values			
Classical	$p = 0.56$	$p = 0.000$	
Wild Bootstrap	$p = 0.68$	$p = 0.024$	
Randomization Inference	$p = 0.74$	$p = 0.082$	
	Panel B: Magnitude		
Election	17	7.1	−10
	(1.7)	(1.1)	(2.0)
	p-values		
	Classical		$p = 0.000$
	Wild Bootstrap		$p = 0.039$
	Randomization Inference		$p = 0.010$
Appointment	23	17	−6.6
	(1.9)	(1.7)	(2.5)
	p-values		
	Classical		$p = 0.010$
	Wild Bootstrap		$p = 0.030$
	Randomization Inference		$p = 0.010$
Difference	−6.0	−9.8	
	(2.5)	(2.0)	
p-values			
Classical	$p = 0.017$	$p = 0.000$	
Wild Bootstrap	$p = 0.19$	$p = 0.047$	
Randomization Inference	$p = 0.23$	$p = 0.11$	

Notes: Standard errors in parentheses.

Table A.4: The impact of distance between a leader's and citizen's types on the probability of choosing a non-selfish policy ($N = 777$).

	(1)	(2)	(3)
Distance	0.0013	0.0011	0.0008
	(0.0007)	(0.0007)	(0.0010)
p-values			
Classical	$p = 0.080$	$p = 0.14$	$p = 0.42$
Wild Bootstrap	$p = 0.20$	$p = 0.24$	$p = 0.56$
Randomization Inference	$p = 0.086$	$p = 0.13$	$p = 0.40$
Election		−0.16	−0.17
		(0.035)	(0.056)
p-values			
Classical		$p = 0.000$	$p = 0.002$
Wild Bootstrap		$p = 0.088$	$p = 0.17$
Randomization Inference		$p = 0.083$	$p = 0.22$
Distance × Election			0.0006
			(0.0015)
p-values			
Classical			$p = 0.68$
Wild Bootstrap			$p = 0.73$
Randomization Inference		$p = 0.083$	$p = 0.80$
Constant	0.57	0.66	0.67
	(0.029)	(0.034)	(0.040)

Notes: Standard errors in parentheses.

Table A.5: Toward whom do leaders move when they move?

	Voter	Losing Candidate
Election	0.62	0.55
($N = 205$)	(0.034)	(0.035)
Appointment	0.61	0.53
($N = 271$)	(0.030)	(0.030)
Election	0.64	0.36
Leader is in between ($N = 70$)	(0.058)	(0.058)
Appointment	0.61	0.39
Leader is in between ($N = 56$)	(0.066)	(0.066)

Notes: Standard errors in parentheses.

Table A.6: The impact of distance between a leader's and citizen's types on the probability of choosing a non-selfish policy.

	Election	Appointment	Difference
Leader is the	0.29	0.33	-0.05
further candidate	(0.087)	(0.060)	(0.11)
	$N = 28$	$N = 63$	
	p-values		
	Classical		$p = 0.65$
	Wild Bootstrap		$p = 0.75$
	Randomization Inference		$p = 0.82$
Leader is the	0.24	0.18	0.06
closer candidate	(0.043)	(0.045)	(0.063)
	$N = 100$	$N = 73$	
	p-values		
	Classical		$p = 0.33$
	Wild Bootstrap		$p = 0.53$
	Randomization Inference		$p = 0.72$
Difference	0.05	0.16	
	(0.093)	(0.074)	
p-values			
Classical	$p = 0.62$	$p = 0.036$	
Wild Bootstrap	$p = 0.70$	$p = 0.23$	
Randomization Inference	$p = 0.81$	$p = 0.047$	

Notes: Standard errors in parentheses. Since we allow for integer amounts, Citizen being in between two candidates is defined as *Leader's type* -1 > *Citizen's type* > *Loser's type* $+1$ or *Leader's type*$+1$ > *Citizen's type* > *Loser's type* -1 so that there is always room for the leader to compromise if he or she wants. Also, *Leader's type* $= 0$ and *Leader's type* $= 100$ are excluded to avoid any movement to favor moving toward the Citizen. z-Values and p-values are based on logistic regression of choosing non-selfish policy on dummy variable indicating the independent variable.

Table A.7: How much do leaders move toward voters (μ) and toward losing candidate (μ')? Average movement relative to initial distance (see paper for exact definition).

	Election	Appointment	Difference
μ	0.46	0.47	-0.005
	(0.030)	(0.028)	(0.041)
	$N = 98$	$N = 125$	
		p-values	
		Classical	$p = 0.90$
		Wild Bootstrap	$p = 0.98$
		Randomization Inference	$p = 1$
μ'	0.43	0.43	0.003
	(0.033)	(0.028)	(0.043)
	$N = 95$	$N = 116$	
		p-values	
		Classical	$p = 0.94$
		Wild Bootstrap	$p = 0.98$
		Randomization Inference	$p = 1$

Notes: Standard errors in parentheses.These values are conditional on moving towards the citizen ($0 < \mu \leq 1$ and $0 < \mu' \leq 1$). $-Valuesand$-values are based on the coefficient of the election dummy variable in OLS regression on a constant and an election dummy variable.

Table A.8: Payoffs.

	Election	Appointment	Difference
Leader	88	80	7.9
	(1.0)	(1.3)	(0.71)
	p-values		
	Classical		$p = 0.000$
	Wild Bootstrap		$p = 0.10$
	Randomization Inference		$p = 0.12$
Losing Candidate	68	69	-1.2
	(1.2)	(1.2)	(1.6)
	p-values		
	Classical		$p = 0.47$
	Wild Bootstrap		$p = 0.61$
	Randomization Inference		$p = 0.60$
Citizen	73	70	2.8
	(1.2)	(1.3)	(1.6)
	p-values		
	Classical		$p = 0.11$
	Wild Bootstrap		$p = 0.19$
	Randomization Inference		$p = 0.17$
Total	228	219	9.4
	(2.2)	(2.5)	(2.4)
	p-values		
	Classical		$p = 0.004$
	Wild Bootstrap		$p = 0.21$
	Randomization Inference		$p = 0.26$

Notes: Standard errors in parentheses. $-Values and$-values are based on the coefficient of the election dummy variable in OLS regression on a constant and an election dummy variable.

B Re-analysis of Data from Original Experiment

The tables in this appendix are the same as those in Drazen and Ozbay (2019), however, we present three different types of hypothesis tests, as discussed in Section 4.2. Additionally, while some specifications in Drazen and Ozbay (2019) used a discrete-choice Logit models, all of these analyses here use OLS.

Table B.2: The impact of being elected on choosing a non-selfish policy ($N = 793$)

Panel A: Fraction Non-Selfish					
Election	0.14	0.14	0.15	0.16	0.16
	(0.033)	(0.033)	(0.033)	(0.036)	(0.036)
p-values					
Classical	$p = 0.000$	$p = 0.000$	$p = 0.000$	$p = 0.000$	$p = 0.000$
Wild Bootstrap	$p = 0.043$	$p = 0.043$	$p = 0.043$	$p = 0.031$	$p = 0.035$
Randomization Inference	$p = 0.059$	$p = 0.057$	$p = 0.056$	$p = 0.053$	$p = 0.057$
Leader's type		0.0002	0.0004	0.0003	0.0004
		(0.0006)	(0.0006)	(0.0006)	(0.0006)
Losing candidate's type			−0.0006	−0.0007	−0.0007
			(0.0006)	(0.0006)	(0.0006)
Citizen's type			−0.0004	−0.0004	−0.0003
			(0.0006)	(0.0006)	(0.0006)
Leader being the closest				−0.045	−0.038
				(0.038)	(0.038)
Period					−0.011
					(0.0028)
Constant	0.25	0.24	0.29	0.31	0.42
	(0.023)	(0.038)	(0.054)	(0.058)	(0.063)
Panel B: Magnitude					
Election	2.4	2.4	2.4	2.8	2.7
	(0.71)	(0.71)	(0.71)	(0.77)	(0.76)
p-values					
Classical	$p = 0.000$	$p = 0.000$	$p = 0.000$	$p = 0.000$	$p = 0.000$
Wild Bootstrap	$p = 0.015$	$p = 0.019$	$p = 0.020$	$p = 0.024$	$p = 0.023$
Randomization Inference	$p = 0.060$	$p = 0.053$	$p = 0.057$	$p = 0.028$	$p = 0.032$
Leader's type		−0.012	−0.010	−0.011	−0.011
		(0.013)	(0.013)	(0.013)	(0.013)
Losing candidate's type			−0.0094	−0.010	−0.0099
			(0.012)	(0.012)	(0.012)
Citizen's type			−0.0041	−0.0046	−0.0007
			(0.013)	(0.013)	(0.013)
Leader being the closest				−1.06	−0.85
				(0.82)	(0.81)
Period					−0.31
					(0.060)
Constant	2.3	2.8	3.4	4.0	7.0
	(0.50)	(0.82)	(1.2)	(1.2)	(1.4)

Notes: Standard errors in parentheses. Leader being the closest is the dummy variable that indicates that the absolute tex difference between the leader's type and the ordinary citizen's type is less than the absolute difference between the losing candidate's type and the citizen's type.

Table B.3: The impact of early vs late periods on choosing a non-selfish policy.

	Early	Late	Difference
	Panel A: Fraction Non-Selfish		
Election	0.45	0.35	−0.10
	(0.035)	(0.034)	(0.049)
	p-values		
	Classical		$p = 0.032$
	Wild Bootstrap		$p = 0.11$
	Randomization Inference		$p = 0.036$
Appointment	0.31	0.20	−0.12
	(0.033)	(0.028)	(0.044)
	p-values		
	Classical		$p = 0.007$
	Wild Bootstrap		$p = 0.30$
	Randomization Inference		$p = 0.008$
Difference	0.14	0.15	
	(0.049)	(0.044)	
p-values			
Classical	$p = 0.005$	$p = 0.001$	
Wild Bootstrap	$p = 0.060$	$p = 0.10$	
Randomization Inference	$p = 0.12$	$p = 0.14$	
	Panel B: Magnitude		
Election	6.2	3.2	−3.0
	(0.97)	(0.65)	(1.2)
	p-values		
	Classical		$p = 0.011$
	Wild Bootstrap		$p = 0.045$
	Randomization Inference		$p = 0.000$
Appointment	3.5	1.0	−2.5
	(0.75)	(0.19)	(0.77)
	p-values		
	Classical		$p = 0.001$
	Wild Bootstrap		$p = 0.051$
	Randomization Inference		$p = 0.000$
Difference	2.6	2.2	
	(1.2)	(0.68)	
p-values			
Classical	$p = 0.032$	$p = 0.001$	
Wild Bootstrap	$p = 0.034$	$p = 0.068$	
Randomization Inference	$p = 0.086$	$p = 0.11$	

Notes: Standard errors in parentheses.

Table B.4: The impact of distance between a leader's and citizen's types on the probability of choosing a non-selfish policy ($N = 793$).

	(1)	(2)	(3)
Distance	0.0000	0.0006	0.0012
	(0.0008)	(0.0008)	(0.0010)
p-values			
Classical	$p = 0.96$	$p = 0.39$	$p = 0.26$
Wild Bootstrap	$p = 0.96$	$p = 0.37$	$p = 0.43$
Randomization Inference	$p = 0.97$	$p = 0.39$	$p = 0.23$
Election		0.15	0.18
		(0.034)	(0.054)
p-values			
Classical		$p = 0.000$	$p = 0.001$
Wild Bootstrap		$p = 0.035$	$p = 0.054$
Randomization Inference		$p = 0.053$	$p = 0.14$
Distance × Election			−0.001
			(0.0015)
p-values			
Classical			$p = 0.47$
Wild Bootstrap			$p = 0.43$
Randomization Inference			$p = 0.46$
Constant	0.32	0.23	0.22
	(0.027)	(0.034)	(0.041)

Notes: Standard errors in parentheses.

Table B.5: Toward whom do leaders move when they move?

	Voter	Losing Candidate
Election	0.77	0.56
$(N = 158)$	(0.034)	(0.040)
Appointment	0.81	0.71
$(N = 100)$	(0.039)	(0.046)
Election	0.77	0.23
Leader is in between $(N = 65)$	(0.053)	(0.053)
Appointment	0.62	0.38
Leader is in between $(N = 29)$	(0.092)	(0.092)

Notes: Standard errors in parentheses.

Table B.6: The impact of distance between a leader's and citizen's types on the probability of choosing a non-selfish policy.

	Election	Appointment	Difference
Leader is the further candidate	0.58	0.23	0.35
	(0.12)	(0.049)	(0.11)
	$N = 19$	$N = 77$	
	p-values		
	Classical		$p = 0.002$
	Wild Bootstrap		$p = 0.17$
	Randomization Inference		$p = 0.031$
Leader is the closer candidate	0.35	0.18	0.17
	(0.046)	(0.055)	(0.077)
	$N = 110$	$N = 50$	
	p-values		
	Classical		$p = 0.032$
	Wild Bootstrap		$p = 0.011$
	Randomization Inference		$p = 0.027$
Difference	0.23	0.05	
	(0.12)	(0.075)	
p-values			
Classical	$p = 0.051$	$p = 0.47$	
Wild Bootstrap	$p = 0.29$	$p = 0.30$	
Randomization Inference	$p = 0.076$	$p = 0.50$	

Notes: Standard errors in parentheses. Since we allow for integer amounts, Citizen being in between two candidates is defined tex as *Leader's type* $-1 >$ *Citizen's type* $>$ *Loser's type* $+1$ or *Leader's type*$+1 >$ *Citizen's type* $>$ *Loser's type* -1 so that there is always room for the leader to compromise if he or she wants. Also, *Leader's type* $= 0$ and *Leader's type* $= 100$ are excluded to avoid any movement to favor moving toward the Citizen. $-Valuesand$-values are based on logistic regression of choosing non-selfish policy on dummy variable indicating the independent variable.

Table B.7: How much do leaders move toward voters (μ) and toward losing candidate (μ')? Average movement relative to initial distance (see paper for exact definition).

	Election	Appointment	Difference
μ	0.38	0.27	0.11
	(0.030)	(0.027)	(0.042)
	$N = 114$	$N = 77$	
		p-values	
		Classical	$p = 0.006$
		Wild Bootstrap	$p = 0.37$
		Randomization Inference	$p = 0.35$
μ'	0.22	0.25	-0.024
	(0.0250)	(0.0244)	(0.036)
	$N = 87$	$N = 68$	
		p-values	
		Classical	$p = 0.50$
		Wild Bootstrap	$p = 0.37$
		Randomization Inference	$p = 0.35$

Notes: Standard errors in parentheses.These values are conditional on moving towards the citizen ($0 < \mu \leq 1$ and $0 < \mu' \leq 1$). $-Valuesand$-values are based on the coefficient of the election dummy variable in OLS regression on a constant and an election dummy variable.

Table B.8: Payoffs.

	Election	Appointment	Difference
Leader	95	98	−2.4
	(0.59)	(0.39)	(0.71)
	p-values		
	Classical		$p = 0.001$
	Wild Bootstrap		$p = 0.015$
	Randomization Inference		$p = 0.060$
Losing Candidate	68	69	−0.61
	(1.1)	(1.2)	(1.7)
	p-values		
	Classical		$p = 0.71$
	Wild Bootstrap		$p = 0.54$
	Randomization Inference		$p = 0.54$
Citizen	77	69	8.2
	(1.1)	(1.1)	(1.6)
	p-values		
	Classical		$p = 0.000$
	Wild Bootstrap		$p = 0.002$
	Randomization Inference		$p = 0.027$
Total	241	236	5.1
	(1.6)	(1.7)	(2.4)
	p-values		
	Classical		$p = 0.031$
	Wild Bootstrap		$p = 0.003$
	Randomization Inference		$p = 0.026$

Notes: Standard errors in parentheses. $-Valuesand$-values are based on the coefficient of the election dummy variable in OLS regression on a constant and an election dummy variable.