NBER WORKING PAPER SERIES

A JOURNAL-BASED REPLICATION OF "BEING CHOSEN TO LEAD"

Allan Drazen
Anna Dreber Almenberg
Erkut Y. Ozbay
Erik Snowberg

Working Paper 26444
http://www.nber.org/papers/w26444

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2019

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 further acknowledge the generous support of the Canada Excellence Research Chairs program, which provided funding for this study. We thank Andrew Proctor and Andreas Born for excellent research assistance, and Magnus Johannessson and Andrew Proctor for comments. Dreber thanks the Jan Wallander and Tom Hedelius Foundation, the Knut and Alice Wallenberg Foundation, and the Austrian Science Fund (FWF, SFB F6) for financial support. Drazen thanks the U.S. National Science Foundation (SES 1534132) for financial support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

A Journal-Based Replication of "Being Chosen to Lead"
Allan Drazen, Anna Dreber Almenberg, Erkut Y. Ozbay, and Erik Snowberg
NBER Working Paper No. 26444
November 2019
JEL No. A11,A14,C18,C92

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 contracts for a replication 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.

Allan Drazen
Department of Economics
University of Maryland
College Park, MD 20742
and NBER
drazen@econ.umd.edu

Anna Dreber Almenberg
Stockholm School of Economics
Sweden
anna.dreber@hhs.se

Erkut Y. Ozbay
Department of Economics
University of Maryland
College Park, MD 20742
ozbay@umd.edu

Erik Snowberg
Division of the Humanities
and Social Sciences 228-77
California Institute of Technology
Pasadena, CA 91125
and University of British Columbia
and also NBER
Erik.Snowberg@ubc.ca

1 Introduction

Recent attempts to replicate experimental findings in economics, and other disciplines, have had a disappointing track record. One high-profile study found that 11 out of 18 economics studies in top publications could be replicated (Camerer et al., 2016). Surprisingly, this was seen as a success, likely because the results compared favorably to a high-profile study of psychology results, where only 35 out of 97 results could be replicated (Open Science Collaboration, 2015).[1] These attempts at replication are sometimes received with hostility, as scholars—whether authors of the original study, or those who try to build on it—may see a basis for their careers being challenged.

More replication seems needed, and there is also a need for it to be done quickly, before false positive (or negative) findings are able to take a prominent place in the literature. This note proposes and executes a proof-of-concept of a novel mechanism for replication: journal-based replication. In this mechanism, a replication attempt is contracted by the journal after a study is accepted for publication, but (ideally) before the actual publication occurs.[2] This can ensure greater confidence in published results, while also ensuring that more experimental results are published, both of which are important for increasing external validity, as discussed in the next Section.

Our own proof-of-concept occurred within the *Journal of Public Economics*, considered a top field journal in economics. The experimental study selected for replication, with the enthusiastic support of the authors (who became co-authors of the current manuscript), was the basis of "Does 'Being Chosen to Lead' Induce Non-selfish Behavior? Experimental Evidence on Reciprocity," by Drazen and Ozbay (2019). That study found that elected representatives are more responsive than appointed representatives to the concerns of their constituents, all else equal. The process of replication is described extensively in Section

[1] 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.

[2] We admit to falling short of this second ideal: although the replication was started before acceptance and completed before publication, we have been unforgivably slow in writing up the outcome.

3, which forms the bulk of this manuscript. The results of the replication, an unanticipated finding that in a Spanish participant population the opposite result obtains—namely appointed representatives are more responsive than those who are elected—is described in Section 4. A discussion of these results, and a summation of the merits of, and potential issues with, journal-based replication concludes this note, in Section 5.

2 Background: Reliability and Replication

Concerns about external validity are endemic to social science research (Banerjee et al., 2017; Christensen and Miguel, 2018). A subset of external validity concerns that seem to be amenable to relatively simple solutions can be broadly labeled *reliability*—roughly encapsulated by the question, "If someone did the same study again, would he or she reach similar conclusions?"[3]

Different research methods call for different approaches to establishing reliability. For observational studies, the standard for reliability in economics is *reproducibility*: obtaining relatively similar results using the original study's data and statistical software code (or exactly the same results when simulated methods are not employed; McCullough and Vinod, 2003; Shachar and Nalebuff, 2004).[4] For experimental studies, the standard for reliability is *replicability*—someone running the same experiment with the same parameter values should obtain similar results.[5] While there is an ongoing debate on how to define "similar results" or whether a study replicates (Gelman and Stern, 2006; Cumming, 2008; Verhagen and Wagenmakers, 2014; Open Science Collaboration, 2015; Simonsohn, 2015; Patil et al., 2016) all versions of replicability require running a very similar experiment again.

[3]This is related to, but somewhat distinct from, *credibility*, which has been used in economics to describe the use of theory-driven statistical methods, usually for the purpose of identifying causal pathways and mechanisms (Imbens, 2010; Deaton, 2010).

[4]There are many instances of surprisingly bad reproducibility. Some of this can be explained by lack of data and code, but problems with reproducibility have been found also with available data and code (Dewald et al., 1986; McCullough and Vinod, 2003; McCullough et al., 2006; Chang and Li, 2018).

[5]This should particularly be the case for treatment-control studies, as level effects may differ across populations. Note that although many social science experiments are executed using computers and software, replication does not require using exactly the same software or code.

A persistent problem for establishing the reliability of experimental results is that there are few incentives to replicate particular studies. A finding that a study replicates has very little chance for publication. On the other hand, a finding that a study does not replicate may result in years of feuding with the original study's authors, nitpicking about particulars of the experiments, and so on. Given those incentives, historically, few studies in the social sciences are directly replicated, and even fewer direct replication attempts are published.[6]

In recent years, mass replications have been carried out by teams of scientists (for example, Open Science Collaboration, 2015; Camerer et al., 2016; Ebersole et al., 2016; Schweinsberg et al., 2016; Camerer et al., 2018; Cova et al., 2018; Klein et al., 2018). The results have lead many to declare a "crisis of replication" in the social sciences (for example, Open Science Collaboration 2015; although see Gilbert et al. 2016; Anderson et al. 2016 for dissenting opinions). There has also been an increase in funding dedicated to producing replications (Baker, 2016). While scholars should be commended for lending their time and energy to these efforts, it is unclear if mass replications present a sustainable model for ensuring the reliability of social science findings. In particular, mass replications are not carried out continuously, leading to long time lags between the publication of a finding and some determination as to its reliability. Moreover, some attempts at mass replication do not even disclose the reliability of particular studies, preferring to focus on aggregate statistics in order to get greater cooperation with the authors of existing studies (Young, 2018). Finally, some scholars believe that emphasizing replication may create incentives to emphasize small differences between a replication attempt and the original study, even when this is not warranted or useful (Hoxby, 2005; Evidence Action, 2015; Gilbert et al., 2016; Fiske, 2016).[7]

The possible unsustainability of mass replications has lead scholars to seek more sustain-

[6]Historically, some studies have been replicated by experimenters trying to extend an existing protocol in new directions. However, it is unknown how many studies were stopped when the original protocol failed to produce the desired results.

[7]It is difficult, if not impossible, to construct an argument against widespread replication on scientific grounds. Thus, we understand these authors to be making an argument that the scant incentives that exist for replication result in a mis-allocation of replication effort. Unfortunately, none of these authors provide proposals for alternative schemes to ensure replication.

able models. Of particular interest here are "registered reports" (see https://cos.io/rr/. In this process, an experiment is accepted for publication *before* it is conducted. A number of journals in the social sciences now allow this sort of submission; the only one in economics is the *Journal of Development Economics*. While the primary aim of registered reports is to reduce publication bias, they may also alter other incentives structures that are associated with less reliable findings. In particular, as there is no need to get a specific result out of the experiment, there are unlikely to be incentives to focus on particular sub-populations, special cases of treatments, and so on, including the use of the many "researcher degrees of freedom" (Leamer, 1983; Simmons et al., 2011; Gelman and Loken, 2013).[8] Despite the fact that replication is not a primary goal of the registered reports framework, as a pro-active step taken by a journal, it is a natural point of reference for journal-based replications.

3 Journal-Based Replication

Although we believe we are the first to conduct a journal-based replication, the idea is quite simple: a journal that is considering whether or not to publish an experiment commissions a replication of that experiment by other experimenters. Despite this apparent simplicity, aspects of the implementation will affect the sustainability of the practice, and its effect on the reliability of experimental research. In this section, we lay out what we believe are the broad decision variables, and our choices on each one, before proceeding to a more detailed description of the process—first from the point of view of the journal, and then from the point of view of the authors whose study was being replicated.

3.1 In Theory

[8] As has been pointed out elsewhere, there may be valid reasons to view statistically significant results as more useful to further research than statistically insignificant results (Brodeur et al., 2016). A related proposal is to post working papers that contain an experimental study, and add a co-author to conduct the replication. The journal submission would then include both the original and replication study (Butera and List, 2017).

There are four broad decisions we believe are of primary importance for the sustainability and usefulness of journal-based replications. They are: whether the replication should be conducted before or after the publication decision; the size and scope of the replication attempt; who should pay for it; and what should be done with the data from the replication. We address each of these in turn, describing both our own decision-making, and our views on what the optimal decisions would likely be going forward.

3.1.1 When Should the Replication be Attempted?

In theory, a journal could decide to publish an article before or after the replication attempt. Presumably, in most cases, if the journal made the publication decision after the replication attempt, that decision would be contingent in some way on the results of the attempt.[9] Whether it is optimal to have the publication decision reached before or after the replication attempt depends on specific beliefs about publication bias, as well as the objectives of the journal (for example, how they weigh the importance of innovative results or those likely to replicate). Making the decision after the replication attempt might also drive some submissions to other journals without replication policies, or lead to the replication being left out of subsequent submissions to other journals, perhaps resulting in additional biases. In our case, due to the exploratory nature of this replication attempt, we felt that publication decision should not be affected by the replication attempt.[10]

3.1.2 Size and Scope of the Replication Attempt:

How replications should be conducted is a matter of some discussion in the literature (see, for example, Anderson et al., 2016; Gilbert et al., 2016; Camerer et al., 2018). Generally, replication sample sizes are larger than those in the studies they seek to replicate in order to have high statistical power to detect the original effect size. However, for a number of

[9]Even if the publication decision was made before the attempt, it could be revised in the unlikely case that the replication attempt uncovered evidence of fraud.

[10]However, we started the replication when publication was not assured (but was extremely likely) in order to ensure that the replication attempt itself would not delay publication of the original manuscript.

reasons, we decided on an "exact" replication that would make no modifications beyond correcting typos and similar errors—including to the sample size. The main reason for this was simplicity. In addition, this approach was informed by our answer to the final question below: "What should be done with the results?" We expected the replication attempt to be successful. However, there is evidence suggesting that reported effect sizes in economics are larger than true effect sizes by about 33% (Ioannidis et al., 2017; Camerer et al., 2018). Given this evidence, and the statistical significance of the original study's effects, we believed the original sample size would likely provide sufficient statistical power. This may not always be the case.

3.1.3 Who Should Pay?

As this was a novel project, it was quite easy to find funding. However, checking the couch cushions is unlikely to provide sufficient funds for a relatively regular use of journal-based replications. We suggest three funding models going forward. First, experimenters could include in their grants a request for funds to cover a replication attempt by a journal. As the cost of lab experiments are usually quite small, this should not be a significant burden for those researchers who can raise grant funding.[11] Second, "open" journals often charge publication fees once an article is accepted for publication. These fees are generally in the \$1,000–\$2,000 range. A typical lab experiment has a broader range of costs, from \$1,000 to \$10,000. While this can be an order of magnitude larger, it is not out of the question to make replication fees a standard part journal publication fees. Third, journals, or the societies that run them, could apply for grants themselves to run journal-based replication pilot programs and to subsidize those experimentalists who would like to have their study replicated but lack the financial wherewithal to do so.

3.1.4 What Should be Done with the Results?

[11] Although lab-based studies are quite inexpensive by granting agency standards, field studies are quite a bit more expensive. Full replications of most field studies are still quite a ways off.

Broadly speaking, many researchers view a replication as either a success—the replication generates an estimate of the main treatment effect that is relatively similar to the prior study (in the same direction and statistically significant from zero)—or a failure—the replication generates an estimate that is closer to (and statistically indistinguishable from) zero. This type of replication indicator was the primary indicator in, for example, Open Science Collaboration (2015) and Camerer et al. (2016, 2018). As already discussed, we anticipated the replication would be a success. In that case, we believed that the original paper would be modified to show the estimates of the main treatment effect from each of the original and replication study, and then pool the data for all subsequent analyses.[12] We anticipated there was some chance the replication would be a failure, in which case we presumed that we would make some note of it in the published paper, and then put together a short note with the rest of the results. However, we neglected a third scenario which is, historically, incredibly unlikely: we might get the opposite effect from that found in the paper. This turned out to be what happened. The next section describes in more detail all the things that went wrong (and right) in our attempt at journal-based replication.

3.2 In Practice

3.2.1 From the Journal's Perspective

Snowberg and Dreber came up with the idea to do a journal-based implementation. After kicking the idea around for a while, Snowberg decided to seek permission from the editors of the *Journal of Public Economics* (where Snowberg was and is a co-editor) to conduct a journal-based replication using discretionary funds that Snowberg had available.

The editors of the journal put three conditions on the attempt:

1. All authors must be senior scholars so that delays relating to the replication effort would not affect anyone's career.

[12] These analyses would, in some sense, be pre-registered.

2. Participation in the journal-based replication scheme must be strictly voluntary: it must be entirely clear to any authors Snowberg approached that turning down his request to participate would have no impact on the acceptance decision of the manuscript.

3. Only one attempt would be allowed as the editors were concerned that more replications might create an unfavorable reputation for the journal resulting from the attempt, which might decrease submissions from experimental economists.

The first condition struck Snowberg and Dreber as wise, and the third as practical. The second is more questionable from a scientific perspective, but reveals an unfortunate aspect of scientific publishing (at least in economics). In particular, journal-based replication would likely increase the perceived reliability of experimental studies in a journal. However, it would also likely result in fewer experimental submissions to the journal, which would have an adverse effect on the journal's reputation. Unfortunately, from a given journal's perspective, the latter cost seems likely to outweigh the benefit of more reliable experimental studies.

Snowberg and Dreber further believed it would be beneficial to make the first attempt with a study that seemed likely, prospectively, to result in a successful replication. This criteria was adopted because Snowberg and Dreber thought this would allow the participants, and future readers, to focus could be on the actual journal-based replication exercise rather than on the specific result. Additionally, Snowberg and Dreber decided the replication attempt should be on an experiment run with relatively standard software, as outsourcing replication of this type of study would be much easier than one that involved, say, responses on paper.

Drazen and Ozbay submitted their paper to the *Journal of Public Economics* in late 2017, and the paper was assigned to Snowberg for handling. After sending the paper out for review, Snowberg asked the editors if this paper would be reasonable for the journal-based replication attempt. The editors agreed that it would be. After reviews of the paper were submitted to the journal, and it became clear that the paper would almost certainly

be accepted, Snowberg asked the authors if they would like to be part of a journal-based replication, and the authors quickly agreed.

Snowberg and Dreber prepared the replication while the paper was still under revision. The lab at the University of Valencia, in Spain was selected as the site for replication.[13] This lab runs experiments for remote experimenters. For Snowberg and Dreber this meant that they could ship the code provided by Drazen and Ozbay to the lab at the University of Valencia and implementation would be taken care of there. IRB approval was obtained from the University of British Columbia, and technicians at the University of Valencia went over the code to make sure it complied with the policies of their lab. During this process, the technicians found certain instructions that they perceived as confusing, which resulted in small changes to wording here and there.

The data from the replication was available before the final manuscript was received. We were initially quite worried about the data itself, as it showed a relatively similar estimate of the main treatment effect, but of the opposite sign. After confirming the data and analysis were correct, Snowberg, at least, had a moment of panic. He was reassured by the editors of the *Journal of Public Economics* that this was quite an interesting outcome, and that it should be up to the original authors what to do with the data. Upon reflection, Snowberg agreed that this was a very interesting result: it indicated that the main effect of the experiment might depend on features of the participant population—for example, culture. That is, the uncommon finding of differently signed results between the original and replication experiment lead Snowberg to believe he had overlooked the possibility that the replication might highlight questions of external, rather than internal, validity.

Snowberg spoke with Drazen and Ozbay; they felt the best thing to do would be to acknowledge the replication result in the main paper, and write up a note containing detailed results. You are reading that note now. They also felt there might be some scope for future research in understanding the (potential) differences between their participant population

[13]This selection was made largely on the basis of familiarity with the setting in Valencia.

and the one in Valencia, Spain.

3.2.2 From the Authors' (Drazen's and Ozbay's) Perspective

When the co-editor (Snowberg) asked us if we were willing to take part in the replication exercise, we happily agreed. Generally, we felt that replication of scientific experiments is very important to test the reliability of the results, as discussed above. Given the importance of replicability, we also agreed that journal-based replication was a good way to try to achieve this general goal, and hence, the strategy was worth trying.

The process was largely as described up front by the co-editor. We sent all relevant material—experimental design, code, description of econometric analysis of results—to the journal, and the co-editor kept us abreast of progress. There was perhaps more communication than would be strictly necessary should in-journal replication become a standard exercise, but given the experimental nature of this initial undertaking, this level of communication was understandable.

Two issues arose during the replication attempt. First, there were some small errors in the analysis code. This was quickly resolved.[14] A second issue concerned which statistical procedures should be applied to lab experiments when the treatment is randomized at the session level. The consensus on how to treat experiments with such a randomization is evolving, and the question came up whether we should revisit our analysis as a consequence. After much deliberation, the co-editor decided that as this was not mentioned by the reviewers, it should not be changed during the replication attempt.

The second issue that arose is part of what we see as a broader concern: whether to revise the analysis in light of differences in results that arise in replication. We agreed with the co-editor that in order for in-journal replication to be successful, a clear line should be drawn between correction of errors in the original experiment and the desire to include

[14]This appears to be a side-benefit of journal-based replication, that the journal actually checks the accuracy of the submitted code. While many journals require the submission of data and code for publication, few journals actually check that the code produces the results presented in the paper (Gertler et al., 2018).

additional results after the fact. While additional results may be of interest in their own right, we felt strongly, as did the co-editor, that they should not be added to an article the journal has already agreed to publish, regardless of the results of a replication exercise (that is, where they did not reveal an error in the original experiment). In our case, we decided, in consultation with the co-editor, to note that the replication exercise led to different results, in a footnote. However, we also decided to neither discuss these differences at any length in the original paper nor to run the experiment a third time, rather refer the reader to a forthcoming note that would explain the differences.

We were happy that our experiment was chosen for this initiative, as we felt it indicated confidence in our experimental technique over and above what would be indicated if the paper was accepted. (The journal making clear that acceptance or rejection of our paper was independent of the outcome of the replication was certainly important in our decision). We felt, perhaps immodestly, that our specific experiment might be a good choice. Why? As outlined in the original paper, in designing the experiment, we tried to control for as many factors as possible in order to argue that our results showed reciprocity of elected leaders to the voter who put them in office. Laboratory experiments in economics strive for the standards of controlled experiments used in the hard sciences in a way not directly possible using data from outside the lab, and that was important to us in this experiment. We thought (more immodesty coming) that we did a reasonably good job in this respect. Rerunning the experiment might support this belief, though as discussed below, perhaps it revealed what we had not controlled for.[15]

[15]There was another reason why we were pleased that our specific experiment would be replicated: curiosity about the behavior of political leaders and what motivates their choices. Hence, it was intriguing to test whether elected leaders are more other-regarding than appointed ones in a different subject population, as well as whether this appears to be reciprocity towards those who elected them.

4 Replication Results

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 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. Like our replication, the experiment in Drazen and Ozbay (2019) had 120 participants across 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.

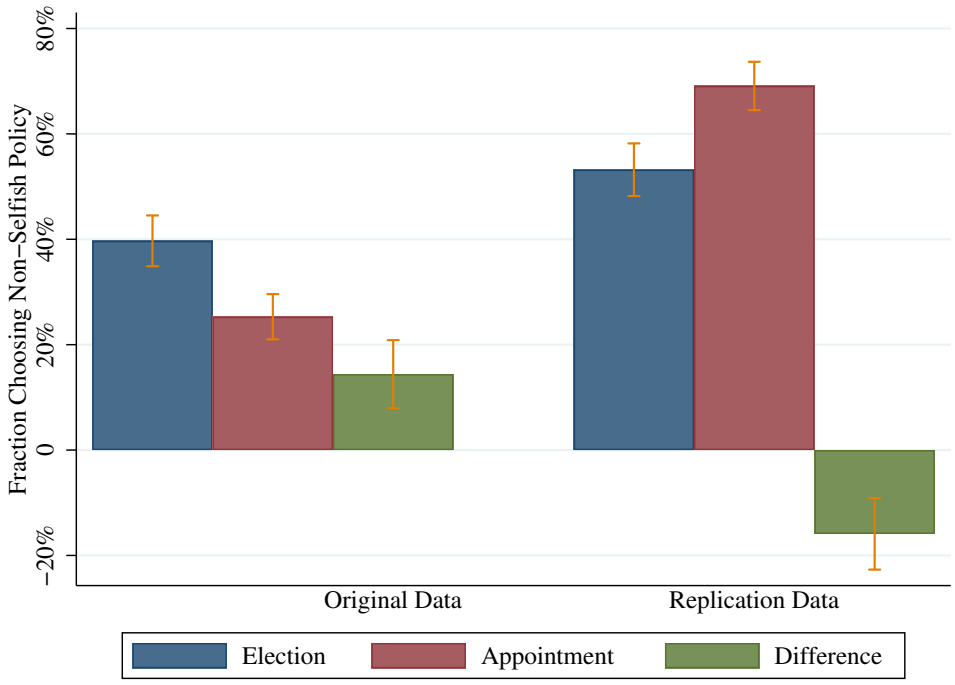
4.2 Main Finding

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. Getting the opposite result is quite rare in replication attempts. We discuss this further in discussion, in Section 5.

It is worth noting other similarities and differences between the replication results, summarized in the Online Appendix, and those found in the tables of the original paper. As seen in Figure 1 and Table 1 of the online appendix, the overall rates of non-selfish behavior are quite a bit higher in both treatments in our replication data.

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.

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



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

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 our main result is different, it is hopefully unsurprising that the results here cannot be explained using those same theories.

5 Discussion

This paper proposes, and shows a proof-of-concept of, a novel mechanism of ensuring replication: journal-based replication. By making publication decisions before a replication is conducted, it reduces the possibility of "file-drawer" problems. Additionally, as the importance of experiments tends to depend on the results of those experiments, it allows editors and reviewers the ability to preview experimental findings (although not replication attempts!) before a publication decision is made.

The biggest concern for this mechanism is 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.[16] Comparing the model of journal-based replication to a model in which replication is achieved through other means (but is still done), 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 can see an argument for these parties bearing the costs of replication. Moreover, if replication were seen as a standard part of the research process, granting agencies would hopefully adjust their funding accordingly, although we suspect this would 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.[17]

The fact that our replication attempt found the opposite result of the original (now-published) study is also worth discussion. There are two plausible interpretations. The first is that both findings are just statistical noise, and that election and appointment do not have any systematic effect on pro-social behavior in this particular experiment. The second 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, Drazen and Ozbay believe that it is this difference that may explain the difference in results. It

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

[17]Dreber and Snowberg's co-author Colin Camerer suggested to us (and in his own grant application to conduct a wide-ranging journal-based replication campaign) that, in equilibria, journals might have two tracks for papers, one where the authors would fund a replication attempt, and one where no such attempt is guaranteed. In the long run, he argued, this 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.

may also be that appointed (versus elected) leaders view their responsibilities towards their constituent populations differently in Spain than in other OECD countries.

Whether one believes the conflicting results are 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 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.

References

Anderson, Christopher J., Štěpán Bahník, Michael Barnett-Cowan, Frank A. Bosco et al., "Response to Comment on 'Estimating the Reproducibility of Psychological Science'," *Science*, 2016, *351* (6277), 1037.

Baker, Monya, "Dutch Agency Launches First Grants Programme Dedicated to Replication," July 20 2016. *Nature*, online at http://dx.doi.org/10.1038/nature.2016.20287.

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.

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," 2017. NBER Working Paper Series # 23335.

Camerer, Colin F. Anna Dreber, 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 et al., "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour*, 2018, *2* (9), 637.

Chang, Andrew and Phillip Li, "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals say 'Often Not'," *Critical Finance Review*, 2018, *7* (1).

Christensen, Garret and Edward Miguel, "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 2018, *56* (3), 920–80.

Cova, Florian, Brent Strickland, Angela Abatista, Aurélien Allard et al., "Estimating the Reproducibility of Experimental Philosophy," *Review of Philosophy and Psychology*, 2018, *10*, 1–36.

Cumming, Geoff, "Replication and p Intervals: p-Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better," *Perspectives on Psychological Science*, 2008, *3* (4), 286–300.

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

Dewald, William G. Jerry G. Thursby, and Richard G. Anderson, "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project," *The American Economic Review*, 1986, *76* (4), 587–603.

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.

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.

Evidence Action, "Worms Win, Kids Lose? Our Statement," 2015. Blog post, available at https://www.evidenceaction.org/worms-win-kids-lose-our-statement/.

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

Fiske, Susan, "Mob Rule or Wisdom of Crowds?," *APS Observer*, October 31 2016.

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.*

_ **and Hal Stern**, "The Difference Between 'Significant' and 'Not Significant' is Not Itself Statistically Significant," *The American Statistician*, 2006, *60* (4), 328–331.

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

Gertler, Paul, Sebastian Galiani, and Mauricio Romero, "How to Make Replication the Norm," *Nature*, 2018, *554* (417), 417–419.

Gilbert, Daniel T. Gary King, Stephen Pettigrew, and Timothy D. Wilson, "Comment on 'Estimating the Reproducibility of Psychological Science'," *Science*, 2016, *351* (6277), 1037–1037.

Hoxby, Caroline, "Competition Among Public Schools: A Reply to Rothstein (2004)," 2005. NBER 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.

Ioannidis, John P.A. Tom D. Stanley, and Hristos Doucouliagos, "The Power of Bias in Economics Research," *The Economic Journal*, 2017, *127* (605), F236–F265.

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.

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

McCullough, Bruce D. and Hrishikesh D. Vinod, "Verifying the Solution from a Nonlinear Solver: A Case Study: Reply," *American Economic Review*, 2003, *93* (3), 873–892.

_ , Kerry Anne McGeary, and Teresa D. Harrison, "Lessons from the JMCB Archive," *Journal of Money, Credit, and Banking*, 2006, *38* (4), 1093–1107.

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

Patil, Prasad, Roger D Peng, and Jeffrey T Leek, "What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science," *Perspectives on Psychological Science*, 2016, *11* (4), 539–544.

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*.

Schweinsberg, Martin, Nikhil Madan, Michelangelo Vianello, S Amy Sommer, Jennifer Jordan, Warren Tierney, Eli Awtrey, Luke Lei Zhu, Daniel Diermeier, Justin E Heinze et al., "The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline," *Journal of Experimental Social Psychology*, 2016, *66*, 55–67.

Shachar, Ron and Barry Nalebuff, "Verifying the Solution from a Nonlinear Solver: A Case Study: Comment," *American Economic Review*, 2004, *94* (1), 382–390.

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, "Accepting the Null: Where to Draw the Line?," 2015. http://datacolada.org/42.

Verhagen, Josine and Eric-Jan Wagenmakers, "Bayesian Tests to Quantify the Result of a Replication Attempt.," *Journal of Experimental Psychology: General*, 2014, *143* (4), 1457.

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

Young, Alwyn, "Consistency Without Inference: Instrumental Variables in Practical Application," 2018. *mimeo.*, London School of Economics.

Online Appendix—Not Intended for Publication

Additional Replication Results and Interpretation

All tables in this online appendix correspond to tables in the paper by Drazen and Ozbay (2019). 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.

Table 1: Do leaders behave non-selfishly?

	Fraction	Magnitude
Election	0.53	12.3
$N = 385$	(0.025)	(1.04)
Appointment	0.69	20.2
$N = 392$	(0.023)	(1.28)
	$z = 4.52$	$t = 4.78$
	$p = 0.000$	$p = 0.000$

Notes: Standard errors in parentheses. z-Values and p-values are based on the significance of the coefficient of the election dummy variable in logistic regression of choosing the non-selfish behavior (column 1) and in OLS regression of the magnitude of non-selfish behavior (column 2) on a constant and election dummy variable.

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

Panel A: Fraction Non-Selfish					
Election	−0.68	−0.68	−0.67	−0.63	−0.70
	(0.15)	(0.15)	(0.15)	(0.15)	(0.16)
Leader's type		0.001	0.001	0.001	0.002
		(0.0026)	(0.0026)	(0.0026)	(0.0027)
Losing candidate's type			0.001	0.001	0.000
			(0.0026)	(0.0026)	(0.0027)
Citizen's type			−0.003	−0.003	−0.001
			(0.0027)	(0.0027)	(0.0028)
Leader being the closest				−0.16	−0.075
				(0.16)	(0.17)
Period					−0.091
					(0.014)
Constant	0.81	0.75	0.84	0.93	1.77
	(0.11)	(0.17)	(0.25)	(0.27)	(0.31)
Log likelihood	−508	−508	−507	−507	−484
Panel B: Magnitude					
Election	−7.9	−8.0	−7.9	−7.0	−7.2
	(1.7)	(1.6)	(1.6)	(1.7)	(1.7)
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	−2.762
				(1.7)	(1.7)
Period					−0.846
					(0.1408)
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 3: The impact of early vs late periods on choosing a non-selfish policy.

Panel A: Fraction Non-Selfish			
Election	0.71	0.35	$z = 6.80$
	(0.033)	(0.035)	$p = 0.000$
Appointment	0.74	0.65	$z = 1.92$
	(0.032)	(0.034)	$p = 0.055$
	$z = 0.58$	$z = 5.65$	
	$p = 0.28$	$p = 0.000$	
Panel B: Magnitude			
Election	17	7.1	$z = 5.16$
	(1.7)	(1.1)	$p = 0.000$
Appointment	23	17	$z = 2.59$
	(1.9)	(1.7)	$p = 0.001$
	$z = 2.38$	$z = 4.78$	
	$p = 0.017$	$p = 0.000$	

Notes: Standard errors in parentheses.

Table 4: The impact of distance between a leader's and citizen's types on the probability of choosing a non-selfish policy.

	(1)	(2)	(3)
Distance	0.006	0.005	0.004
	(0.0032)	(0.0032)	(0.0046)
Election		-0.66	-0.72
		(0.15)	(0.24)
Distance × Election			0.002
			(0.0065)
Constant	0.29	0.66	0.69
	(0.12)	(0.15)	(0.18)
Observations	777	777	777
Log likelihood	−517	−507	−507

Notes: Standard errors in parentheses.

Table 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	(0.058)	(0.058)
N = 70		
Appointment	0.61	0.39
Leader is in between	(0.066)	(0.066)
N = 56		

Notes: Standard errors in parentheses.

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

	Election	Appointment	
Leader is the	0.29	0.33	$z = 0.45$
further candidate	(0.087)	(0.060)	$p = 0.65$
Leader is the	0.24	0.18	$z = 1.03$
closer candidate	(0.043)	(0.045)	$p = 0.30$
	$z = 0.55$	$z = 2.29$	
	$p = 0.58$	$p = 0.022$	

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 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	
μ	0.46	0.47	$z = 0.01$
	(0.030)	(0.028)	$p = 0.99$
	N=98	N=125	
μ'	0.43	0.43	$z = 0.01$
	(0.033)	(0.028)	$p = 0.99$
	N=95	N=116	

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

Table 8: Payoffs.

	Election	Appointment	
Leader	88	80	$z = 4.78$
	(1.0)	(1.3)	$p = 0.000$
Losing Candidate	68	69	$z = 0.72$
	(1.2)	(1.2)	$p = 0.47$
Citizen	73	70	$z = 1.61$
	(1.2)	(1.3)	$p = 0.11$
Total	228	219	$z = 2.87$
	(2.2)	(2.5)	$p = 0.004$

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