Good and Bad Replications in Political Science: How Replicators and Original Authors (Should) Talk to Each Other

Nicole Janz, *University of Nottingham, nicole.janz@nottingham.ac.uk*

Jeremy Freese, *Stanford University, jfreese@stanford.edu*

Introduction

Reproducibility is seen as a cornerstone of scientific research. Even though the value of replications is more and more recognized in political science (e.g. Carsey 2014, Nyhan 2015, Elman, Kapiszewski and Lupia 2018), a cursory look at journal archives shows that the publication of replication studies is not yet common. This is surprising, especially when compared to the trend towards large replication initiatives in other fields such as psychology (Open Science Collaboration 2015) or cancer research (Errington et al. 2014). Replication activities are also becoming more common within economics (Christensen and Miguel 2018), and the ReplicationWiki database collects published replications in economics to make it easier to find verifications of published work (Höffler 2017).

When replications in political science are conducted, they may be often land in the file drawer anyway because they may not be seen as novel enough to be published (Carsey 2014). Especially replications conducted by graduate students remain an untapped resource (Janz, Werfel, and Wykstra 2014, Janz 2016). Those replication studies that have been published are hard to find; they appear to be framed as large extensions, almost like a new research article (e.g. Miller et al. 2001, Dai 2002). Only very recently has "replication" made into the title of published articles (Coppock 2018, Busby and Druckman 2018).[1]

One way to promote the conduction and publication of more replication studies is a change in research culture so that honest mistakes are normalised and avoid embarrassment for original authors (Elman, Kapiszewski and Lupia 2018). Our essay shows how this can be implemented in

[1] Notably, Coppock's (2018) replication of an earlier study confirms these findings while extending the experiement to provide additional insights: "The bottom line for the substantive results of the Butler and Nickerson (2011) experiment has not changed—if anything, the magnitude of the total causal effect has nearly doubled." (Coppock 2018: 167).

practice. We discuss 'good and bad' replications in political science and show how well-conducted and constructive replications can become publishable (for the replicators), and that there is nothing to fear (for the original authors).

Replication versus Duplication

Duplication is the process of "running the same analyses on the same data to get to the same result" (King 1995: 451). In practical terms, a duplication can involve using the replication data set and software code provided by the original author, or, for observational studies based on administrative and publicly available data, the re-collection of the same data from the primary sources, e.g. the World Bank. Duplications are very useful as student assignments to teach statistics (King 1995, Carsey 2014, Janz 2016). Duplications are also important outside of the classroom because if any errors occurred that reverse findings, this should be publicly known (in a way that assumes the good faith of investigators and recognizes that mistakes happen). The reproduction of existing results is important to strengthen our certainty of present results so that we can meaningfully build on such work (King 1995).

Replication is the process by which a study's hypotheses and findings are re-examined using different data or different methods, or both. Replications address the question of whether the study's results might be due to chance, to arbitrary or suboptimal methodological decisions, or to narrow contexts in which data were collected (Freese & Peterson 2018). In this way, replication studies assess the robustness of a result (Herrnson 1995) and seek to advance knowledge from existing research.

The difference between duplication and replication comes into play in how we present deviations from the original study. The reason for a failed duplication is knowable, even if it can be maddeningly difficult to figure out. It often implies a mistake, even if the mistake might be the duplicator's, rather than any errors in the originating study. The most superficial duplication problems--executing the same code with the same software on the same data and getting different results--are often readily preventable by authors of original studies following best practices for reproducibility. For example, any code that employs random numbers needs to provide information about the random seed; information about the version of data needs to be provided if a dataset must be obtained from a third-party instead of as part of the replication package; and the process of creating manuscript tables from results should be automated as much as possible to avoid transcription errors (Christensen, Freese, and Miguel forthcoming). Some journals (e.g. *Political Science Research and Methods* or the *American Journal of Political Science*) try to prevent such issues by re-running the original analysis based on the replication materials before publication, and many scholars recommend that authors duplicate their own

findings, or ask a colleague or student to do this, even before journal submission (Markowetz 2015).

When results of a replication study diverge from those of an originating study, interpretation is far more complicated than for a duplication. Whenever new statistics and new data are involved, there is some possibility that divergent results are simply due to chance. Often divergent results are plausibly due to any number of differences, e.g. the choice of operationalisation and measurement of variables, updated data sets, statistical methods choices, larger contexts, omitted variables, missing data handling, or newly updated and revised observational data.

With new data, reasons for divergent results are often mysterious without other studies and still more data, and even then commonly elude decisive explanation. Divergent findings from a replication study indicate that findings from the originating study are not as robust as one might have thought. If the replication study can point to reasons why it is better than the originating study (e.g. better data, measures, or methods), a replication study might be viewed as more authoritative than the original study. King's (1995: 445) advises that replicators "improve on the data or methodology." Prioritizing the replication over the original article simply based on the chronology of publication is, of course, debatable (Gelman 2018), and original authors tend to comment negatively on the quality of replications, as we discuss below. The connotations of "failed replication" is almost always too strong for what divergent results indicate, and the wording used by replicators is a delicate matter.

The issues at stake in a replication study usually involve limitations that exist in pretty much any study published at a given point in time. Many issues are often already discussed in the robustness checks section of the original article, but such space is often limited, and new methods, data and approaches may have been developed after publication. In a way, a replication reflects typical research challenges in particular fields, and ideally, should be framed as such. In particular, even when a well-conducted "failed" replication causes us to reassess a previous finding, we should not automatically interpret it as discrediting an original author's diligence, expertise or choice of methods. Coming to different results when you use new data or methods does not necessarily indicate that the original author's work was faulty. Those publishing replication studies with divergent results should endeavor to provide the clearest explanation of what they can determine about why the results differ.

"Replication chains": How replicators and original authors talk to each other

The few replications published in political science have been relatively harsh in their wordings, and have provoked what has been labelled `replication chains'[2] from original study, to subsequent replication, and a further commentary by the original author on the replication. Most original authors have defended their earlier paper by stating that the replication was fundamentally flawed.

For example, a highly cited randomized field experiment published in the *American Political Science Review* showed that voter turnout was increased substantially by personal canvassing but not by telephone calls, calling into question millions of dollars spent on phone canvassing during for elections (Gerber and Green 2000). A replication study later found that "Gerber and Green's negative finding is caused by inadvertent deviations from their stated experimental protocol" and pointed to "systematic patterns of implementation errors" (Imai 2005: 283). The replication study used the term 'errors' over 15 times. The original authors replied to the replication and stated that the replication "is shown to contain statistical, computational, and reporting errors that invalidate its conclusions" (Gerber and Green 2005: 301), referring to "Professor Imai" repeatedly, e.g. when stating "none of the key substantive or methodological claims of Professor Imai's essay survives scrutiny" (pg. 301). This replication chain shows that the language is--while being professional--still harsh between original authors and replicators. It would be much more useful to refer to the study itself rather than the author; it might also be more constructive to discuss "deviant findings" or "discrepancies" rather than errors.

A different replication chain uses slightly more constructive wording, at least in some parts. An original article on citizens' political tolerance (Peffley, Knigge, and Hurwitz 2001) was replicated with the words "We regret to say that we found some significant differences when attempting to replicate the … results." (Miller et al. 2001: 407). The replicators stated two main reasons for differences in the results, and showed how they tried to double-check if they themselves were at fault ("we wanted to give the original analyses the benefit of the doubt", pg. 408). The replication concluded that the original authors "made a simple coding mistake" and that "these analysis errors are not significant enough to dismiss this article totally, they are troublesome" (pg. 409). The language of the replication was relatively measured, except for where the replicators note that they are worried about the errors "particularly given the professional standing of the senior authors" (pg. 409), a comment which we find certainly unnecessary. The original authors then stated that the replication is "based on a fundamentally flawed analysis", that their their criticisms "suffer from a limited understanding of existing theory and research in the area", and that the alarm raised by the replicators is "seriously exaggerated" (Peffley et al. 2001: 421-422). In addition, they turned it around and tried to duplicate the replication's findings, without success, concluding that "the differences between

[2] See several blog posts on replication chains on the *Political Science Replication Blog* available at https://politicalsciencereplication.wordpress.com/category/replication-chains/ (Accessed January 2, 2019)

our findings and Miller et al.'s replication are trivial and not demonstrative of any noteworthy flaws in our original analysis." (pg. 422).

In these replication chains, we noticed that the replicators as well as the original authors' used a detailed comparison of differences of results, but their substantial interpretation of these differences, and the categorisation of these as strong or trivial, and the replication as failed or not, varied. The replicators always described these differences as large and worrisome, while the original authors replied that--even if minor errors occurred--the main findings remained the same. None of original authors contested the importance of replication as such, but they implied that the replicators lacked understanding of the topic at hand. This is not surprising. There is seldom complete agreement in scholarly research and 'the truth' probably does not exist. Of course, the disagreement and harsh language could have to do with the agenda of replicators (to get published) and the original authors (to confirm their earlier findings).

These issues surrounding replication and disagreements about 'who is right' may seem off-putting. However, there are many good ways to conduct replications so that honest mistakes--or different views about approaches--are normalised. Our golden rule is: *replicate others as you would like to be replicated yourself*! The next part shows suggestions on how this can be implemented in practice. Constructive replications are publishable, and if conducted with nuance, there is nothing to fear for the original authors.

A good replication

The expert knowledge and professionalism of the replicator should be reflected in their careful and transparent planning of a replication study as well as in their professional language towards the original study. We propose two main ways to conduct a well-conducted, constructive and publishable replication: First, your study should be carefully and transparently planned. In particular, make very clear if the replication aims to conduct a duplication or replication study, or both, and have a clear list for yourself of what differences between the results you expect when you use new data or methods. Second, a good replication study should also use sensitive and professional wording when referring to the original study to avoid embarrassment for the original author(s). We emphasize that binary judgments such as 'failed replication' can do more harm than good. We discuss several examples for these two main features and show pitfalls to avoid.

We start with an example: The recently published study "Football and Public Opinion: A Partial Replication and Extension" (Busby and Druckman 2018) focused on existing literature asking if irrelevant events, such as the weather, sporting events, and random lotteries, can influence political behaviour or attitudes. Before setting out to replicate one of the previous studies on the topic, the authors set the tone:

> "To be clear, this is not a critique of existing papers, which faithfully report careful studies that establish the existence of irrelevant event effects (i.e., researchers did not actively set up studies most likely to produce effects). Rather, replication with a different event, sample, and time is a way to move the literature forward to assess robustness and the conditions under which irrelevant event effects occur." (pg. 5).

The replicators' goal was not to criticize previous work or hunt for errors, but to move the literature forward. They also acknowledged that the original authors did not in fact claim generalisability. The replication study goes on to same experimental design and procedure as the original study, what extensions are planned, and finally describes how their results partially differ from the previous result. The authors do not use a binary judgment such as that the previous study failed to replicate, but they state clearly which results replicate, and which did not, writing that "we replicated one such effect … we failed to replicate the effect for …". (pg. 7-8). The replicators also refrain from claiming to provide the final answer to the question, emphasizing that their result should "not be taken as definitive evidence that the extant literature over-states the extent of irrelevant events; yet, it serves as a (cautionary) prompt to the next generation of work." (pg. 8).

In summary, what we can learn from this replication study is that it clearly states what differences are expected based on the duplication and extension, and it expresses its views about the replicability of the previous results with nuance rather than a binary, 'alarming' judgment. There is only one shortcoming of this example: Busby and Druckman (2018) were replicating their own study (Busby, Druckman, and Fredendall 2017). Does this mean that this example is irrelevant? On the contrary, this replication study, which is almost a form of partial self-correction, serves exactly the point we are making. Our golden rule is: replicate others as you would like to be replicated yourself! Imagine that you look back on one of your previous studies and you feel that there is something missing, or a follow-up and cross-check is in order to advance the literature on the topic. Then plan and write your replication (of another study) exactly as you would have done it for yourself. Here are additional suggestions on how this can work:[3]

Careful and transparent planning of a replication

[3] Some of these can also be found in the submission policy of the *Political Science Replication Initiative* (PSRI), see http://projects.iq.harvard.edu/psreplication (accessed January 2, 2019).

1. Make clear: Are you conducting a replication or duplication? How 'far' must your results deviate from the original work before claiming that some effects could not be confirmed? Take into account that different data and measurements (replication) will naturally yield different results, rather than expecting that the original study is flawed.

2. Be transparent and reproducible: Why have you chosen the original study for replication? Could selection bias be at work? Also, your study must have an extensive and clear methods part, and potentially a supplement, to give all details on how the new findings were produced. Ideally, a replication would be pre-registered to eliminate reporting flexibility (Zigerell 2017). There is evidence from experimental psychology that a publication bias for replication studies (towards failed replications) exists (Francis 2012), and transparency can avoid later accusations of p-hacking or deliberate error hunting. Have someone crosscheck your replication before journal submission.

3. Be an expert: Engage deeply with the substantive literature, not just the original study, to ensure that the interpretation of differences between original and replication is thorough and acceptable to authors in the field. Ideally, extensions should be from a clear theoretical argument or methodological critique (King 2006).

Rhetorical sensitivity towards original study

1. Avoid binary judgments: Explain replication results constructively and be fair to the original author(s), while still pointing out issues with robustness and improving research on the topic. Present your results step by step, and show exactly which effects were replicated, which were not, and interpret why this might be the case (King 2006). Divergent interpretations of the meaning of "failed" replications is all over the history of replication in science, so we recommend avoiding that the replication of a study has "failed," except maybe when you uncover a case of extreme scientific misconduct (see Broockman, Kalla and Aronow 2015).

2. Don't make it personal: Use professional, courteous and collaborative (as opposed to confrontational) language. Always try to talk about the study, not the author, to make it less personal. Do not imply that original authors claimed generalisability when in fact they point to certain limitations themselves. Make clear what the positive contribution of the original article is--after all, you would not have chosen it if it the study to be re-examined was not crucial to the field. Honest mistakes are human, and they should not be used to discredit an original author. When using social media to promote your replication, be constructive. Gary King (2006) advises that a copy of the replication

should first be sent to the original author for comments and feedback. You may want to leave this decision to the journal editors.

3. Look forward, not backwards: How do your findings fit with the rest of the literature, and who can benefit what from your replication? Avoid the notion that your judgment on the original study may be 'final'. Discuss what the literature learns from the replication. Gary King advises that a copy of the paper should first be sent to the original author by his students, who can respond to the critique and comment on possible failed replications. In psychology, there have been calls to involve the original authors even earlier, as part of the planning or pre-registration process (Kahneman 2014, Nosek & Lakens 2014).

Conclusion

Ideally replication studies in political science would be much more routine and much less emotionally freighted. Fundamentally, replications are research, which contribute to our collective effort to build knowledge and improve on past research. We need to develop a research culture in which authors view efforts to replicate their work as a sign of interest and importance, rather than as something to be feared.

We have focused here on what replicators can do to provide the most constructive replications. Of course, journals have an important role to reply in cultivating a research culture in which conductive constructive replications are rewarded. We think that it is important that journals not invert the publication bias of original studies toward positive findings by only publishing "failed" replications, but instead focus on the quality of replications. Also, journals can provide guidelines for good replications analogous to what some provide about reproducibility.

Journals can also help make replications more constructive by working to make sure that original studies provide as much information to facilitate replication as possible. Guidelines that journal can adopt to promote sharing have been offered (e.g., TOP guidelines; Nosek et al. 2015). When both replicators and original authors adhere to the highest transparency standards and communicate their concerns professionally, replications will become more acceptable and are a welcome way to add knowledge to the field.

References

Bell, M. S., & Miller, N. L. (2015). Questioning the effect of nuclear weapons on conflict. *Journal of Conflict Resolution*, *59*(1), 74-92

Broockman, D., Kalla, J. & Aronow, P. (2015). "Irregularities in LaCour (2014)." Available at: http://web.stanford.edu/~dbroock/broockman_kalla_aronow_lg_ irregularities.pdf (accessed January 2, 2019).

Busby, E. C., Druckman, J. N., and Fredendall, A. (2017). "The Political Relevance of Irrelevant Events." *The Journal of Politics*, 79(1), 346–50.

Busby, E. C., & Druckman, J. N. (2018). Football and Public Opinion: A Partial Replication and Extension. *Journal of Experimental Political Science*, *5*(1), 4-10

Carsey, T. M. (2014). Making DA-RT a reality. *PS: Political Science & Politics*, *47*(1), 72-77

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920-80

Christensen, G., Freese, J. & Miguel, E. (forthcoming). Transparent and reproducible social science research: How to do open science. University of California Press.

Coppock, A. (2018). Generalizing from survey experiments conducted on mechanical Turk: A replication approach. *Political Science Research and Methods*, 1-16

Dai, X. (2002). Political regimes and international trade: The democratic difference revisited. *American Political Science Review*, *96*(1), 159-165

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). Science forum: An open investigation of the reproducibility of cancer biology research. *Elife*, *3*, e04333

Freese, J., & Peterson, D. (2017). Replication in social science. *Annual Review of Sociology*, *43*, 147-165.

Gelman, A. (2018). Don't characterize replications as successes or failures. Behavioral and Brain Sciences, 41, E128. doi:10.1017/S0140525X18000638

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

Gerber, A. S., & Green, D. P. (2005). Correction to Gerber and Green (2000), replication of disputed findings, and reply to Imai (2005). *American Political Science Review*, *99*(2), 301-313

Höffler, J. H. (2017). Replication and economics journal policies. *American Economic Review*, *107*(5), 52-55

Janz, N., Werfel, S., & Wykstra, S. (2014). “Replication in Political Science Graduate Courses: an Untapped Resource?” *Monkey Cage - The Washington Post*. February 12, 2014

Janz, N. (2014). “ ‘Replication Bullying:’ Who replicates the replicators?” *Political Science Replication Blog*. May 25, 2014. Available at: https://politicalsciencereplication.wordpress.com/2014/05/25/replication-bullying-who-replicates-the-replicators/ (accessed January 2, 2019).

Janz, N. (2016). Bringing the gold standard into the classroom: replication in university teaching. *International Studies Perspectives*, *17*(4), 392-407

Imai, K. (2005). Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. *American Political Science Review*, *99*(2), 283-300

Kahneman, D. (2014). A new etiquette for replication. *Social Psychology, 45*(4), 310-311.

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 444-452

King, G. (2006). Publication, publication. *PS: Political Science & Politics*, *39*(1), 119-125

Markowetz, F. (2015). Five selfish reasons to work reproducibly. *Genome biology*, *16*(1), 274

Miller, A., Wynn, T., Ullrich, P., & Marti, M. (2001). Concept and measurement artifact in multiple values and value conflict models. *Political Research Quarterly*, *54*(2), 407-419

Nosek, B. (2014). “Replications of Important Results in Social Psychology. Special Issue for Social Psychology. Context and Correspondence for one commentary”. Available at: https://docs.google.com/document/d/1ew7X0RaClU5_Ev4Ns3Uyn0I7PmjzP_Z1wKlnza_3Fe0/edit (Accessed January 2, 2019).

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*(3), 137-141.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422-1425.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716

Peffley, M., Knigge, P., & Hurwitz, J. (2001). A multiple values model of political tolerance. *Political Research Quarterly*, *54*(2), 379-406

Peffley, M., Knigge, P., & Hurwitz, J. (2001). A Reply to Miller et al.: Replication Made Simple. *Political Research Quarterly, 54*(2), 421-429