Replications in Context: A Framework for Evaluating New Methods in Quantitative Political Science

Jeffrey J. Harden[1], Anand E. Sokhey[2] and Hannah Wilson[3]

[1] *Department of Political Science, University of Notre Dame, 2055 Jenkins Nanovic Halls, Notre Dame, IN 46556, USA.*
Email: jeff.harden@nd.edu
[2] *Department of Political Science, University of Colorado Boulder, 333 UCB, Boulder, CO 80309, USA.*
Email: anand.sokhey@colorado.edu
[3] *Department of Political Science, University of Notre Dame, 2060 Jenkins Nanovic Halls, Notre Dame, IN 46556, USA.*
Email: hwilson2@nd.edu

Keywords: replication, quantitative methods, researcher degrees of freedom

1 Introduction

Over the past several decades, the subfield of political methodology has facilitated considerable growth and diversification in the set of tools that political scientists use to conduct quantitative research. This expansion—which includes the development of new methods and the importation of methods from other fields—has produced enormous benefits for the discipline. However, an extensive set of tools also necessitates guidance from methodologists regarding where in the methodological landscape applied researchers should invest their finite time and effort. Methodologists often meet this demand, in part, by using replication analyses to demonstrate that a new method holds empirical consequences for researchers' substantive conclusions. However, the dominant approach to conducting these replications can suffer from many of the same problems that concern applied researchers, such as selection bias and unrepresentative samples of data. To address this issue, we propose an alternative framework for the evaluation of new methods.

Methodologists often justify the utility of a new method by replicating past work with a "proof of concept" standard. This approach approximates the practice of most-likely case selection (e.g., Gerring and Cojocaru 2016) for replication studies. The methodologist initially believes that the new method is "better" than an existing method (at least under some conditions), then looks for supporting evidence from at least one replication of published work (i.e., a "crucial case") in which the two sets of results yield different substantive conclusions. We contend that this practice—which only demonstrates that a new method *can be* consequential—is problematic for several reasons. First, it encourages selection bias in replication analyses—methodologists have strong incentive to sift through replication data until they find one or two crucial cases in which the existing and new methods point in different directions. Second, replicating only a few studies provides limited generalizability with respect to the method's potential impact on the research community. Finally, the typical starting point—that the new method is substantively different from the existing method—is more sensible as an alternative hypothesis, not a null.

Political Analysis (2019)
vol. 27:119–125
DOI: 10.1017/pan.2018.54

Corresponding author
Jeffrey J. Harden

Edited by
Jeff Gill

Authors' note: The example presented here was documented in a preanalysis plan deposited at the Political Science Registered Studies Dataverse (doi:10.7910/DVN/J7HFRX) prior to execution. All replication materials are available at the *Political Analysis* Dataverse (Harden, Sokhey, and Wilson 2018). Author names appear in alphabetical order. We thank Justin Kirkland and Carlisle Rainey for helpful comments.

Table 1. Meta-analysis of replications in methods articles, 2008–2018.

Journal	Articles	Mean replications	Justify selection	New method's effect on results			
				Weaker	Stronger	Mixed	Same
APSR	5	2.40	60%	17%	8%	25%	50%
AJPS	24	1.75	30%	31%	19%	31%	19%
PA	49	2.30	37%	43%	11%	20%	25%
Total	78	2.20	36%	38%	13%	23%	25%

These issues may ultimately produce an overconfident view of a new method's utility to substantive research. Addressing this problem of overconfidence is important because the introduction of new methods is not costless. New methods improve substantive research, but they also naturally increase researcher degrees of freedom, magnifying the risk of reporting false positive results. Moreover, a larger methodological portfolio requires applied researchers to keep investing in methods training, perhaps at the expense of developing substantive expertise. Accordingly, we propose the use of a stricter standard in evaluating the practical utility of new methods.

Our alternative framework achieves this objective with several steps: (1) preregistering a replication plan, (2) collecting a representative sample of replication studies, (3) presenting *distributions* of differences between the existing and new methods, and, if feasible, (4) employing Bayesian inference to test the null hypothesis that the new method is not substantively different, on average, from the existing method. In short, our approach encourages methodologists to place replications in context. We illustrate it with a preregistered example that compares two estimators of the Cox proportional hazards model. We ultimately conclude that our approach sets a more rigorous standard for assessing a new method's practical utility, which complements the current trends toward greater transparency in the research process and ease of access to replication data.

2 Replication in Political Methodology

Replication has long been a topic of discussion in political science (e.g., King 1995). These conversations have recently produced changes in disciplinary norms, such as the adoption by several journals of new guidelines and standards regarding access to replication data. Scholarship on replication addresses several important aspects, including transparency standards, workflow, the assignment of responsibility and credit, publication bias in replication studies, and others (e.g., Carsey 2014; Ishiyama 2014; Berinsky, Druckman, and Yamamoto 2018). However, one topic that has received less attention in this conversation is the manner in which methodologists employ replication. In contrast to the typical substantive replication—in which an empirical claim is scrutinized in a new context—the typical methodological replication uses existing work as a testing ground to highlight a new method's value. Methods articles present the statistical case for an innovation through proofs, simulations, and other evidence. Then, the goal of the replication analysis is to offer proof of concept that the method actually matters for applied researchers. Methodological replications often accomplish this goal by showing that the new method alters the substantive conclusions reported in at least one or two published articles.

To illustrate this point, we conducted a meta-analysis of methods articles appearing in three journals—*American Political Science Review*, *American Journal of Political Science*, and *Political Analysis*—during the period 2008–2018. Our search yielded 78 articles that included 169 unique replication studies. We documented several pieces of information from the articles: (1) a count of replication studies, (2) whether or not the article justified the process of selecting replication studies, and (3) the new method's effect on the original studies' substantive conclusions. Table 1 summarizes our results.

Several patterns emerge from our meta-analysis. First, we confirm that the typical methods article replicates a small number of studies (about two, on average). Additionally, it is uncommon (though not extremely rare) to justify the sample of replication studies. We also find evidence consistent with a selection bias problem. Approximately three-fifths of the replications in our sample of articles show weaker results (defined as all hypotheses having weaker support), or mixed results (some hypotheses with weaker results). In fact, we find more reports of complete reversals of the original studies' findings (14%) than reports of stronger results (13%). In short, our meta-analysis highlights the potential problems with methodological replication analyses discussed above.

3 An Alternative Framework

These problems motivate our alternative framework for assessing new methods, which we outline briefly here (see the Supplementary Appendix for a complete description). The framework is comprised of two specific modes of conducting replications, both of which facilitate more comprehensive evaluation of a new method compared to the proof of concept approach. The first mode, which we refer to as a *test of concept replication analysis* (TCRA), is similar to the current approach in that the methodologist uses the replications as illustrative examples of the new method. However, in the TCRA mode the examples are publicly documented prior to conducting the replications. The second mode—the *full inference replication analysis* (FIRA)—goes even further: it treats empirical evaluation of a new method as an inference problem, similar to the manner in which applied researchers test substantive hypotheses.

Both modes begin with a preanalysis plan (e.g., Monogan 2013), which directly addresses the selection bias problem. Preregistering a methodological replication analysis emphasizes the replication context by establishing which studies are relevant in advance of looking at the data. We propose three key elements for inclusion in the preanalysis plan: (1) a definition of the population of studies relevant to the new method, (2) a description of a target sample of (ideally, many) studies drawn from this population for actual replication, and (3) a description of the specific replication quantities of interest (RQI) that will be computed as part of the replication process to measure the differences between the existing and new methods.

Once the preanalysis plan is deposited, the next step is to conduct the replications. If the methodologist chooses the TCRA mode, comparing the methods under study consists of providing descriptive information about the RQI via summary statistics and graphs. Interpretation then focuses on what can be learned within the sample. Conducting the FIRA mode goes one step further, and requires a theory of inference. Standard frequentist hypothesis testing is an option, but it usually requires distributional assumptions that may be suspect if the sample size is small (i.e., few available replication studies). Additionally, frequentist hypothesis testing forces reliance on the language of statistical significance to measure the new method's value. We contend that this approach is too restrictive (Gill 1999).

Our framework addresses these issues by bootstrapping from a Bayesian perspective, with the goal of constructing a posterior distribution for the mean of the RQI.[1] Classical and Bayesian bootstrapping are essentially equivalent in practice. However, the Bayesian approach performs better in small samples (Rubin 1981) and produces a proper posterior that facilitates more useful inferences. We advise formally assessing substantive differences between the methods under study via Bayesian highest posterior density (HPD) intervals. Specifically, Rainey's (2014) method of evaluating hypotheses can be used to determine if large differences in the RQI are plausible. A key aspect of this process involves choosing a cutoff value, m, that defines the smallest substantively meaningful average RQI. In other words, the methodologist should determine how

1 Bootstrapping is also useful because it gives the methodologist flexibility to handle a variety of types of correlation between data points, such as clustering.

different, on average, the existing and new methods must be (according to the RQI) such that it would be inadvisable for an applied researcher to ignore the new method as a possible empirical strategy.

4 Replicating Replications

We illustrate this framework with an example from a methodological article that employs replication.[2] Bednarski (1993) proposes a robust estimator of the Cox proportional hazards model that downweights outliers to reduce coefficient bias stemming from specification problems. Desmarais and Harden (2012, hereafter DH) compare this estimator to the standard partial likelihood estimation method (Cox 1975), and develop the cross-validated median fit (CVMF) test for empirically choosing between the two methods. Here we treat the robust estimator as the new method and the partial likelihood approach as the existing method. DH conduct this same comparison, reporting results from five replication studies. They emphasize two findings: (1) their CVMF test selects the robust method as the better-fitting estimator in three of the five replication studies, and (2) substantive conclusions can change when the robust estimator is used.

In what follows we employ our evaluation approach to replicate DH's replication analysis. We defined the population as any political science study that employs the Cox model, then drew a target sample of 24 studies from several journals in the discipline. We successfully replicated the standard estimator results from one "main" model specification *and* executed the robust method and CVMF test for 11 of these studies. We then added DH's five replications for a total of 16. We evaluate the practical utility of the robust estimator with two RQI. First, at the study level ($n = 16$) we estimate the proportion of studies for which DH's CVMF test selects the robust estimator. Our preanalysis plan states that an estimate of 25% or more would constitute favorable evidence for the practical utility of the robust estimator. Second, we measure divergence in results at the coefficient level ($n = 209$) by computing the ratio of the absolute coefficient estimates from each estimator: $\frac{|\widehat{\beta_{\text{robust}}}|}{|\widehat{\beta_{\text{standard}}}|}$. Values greater (less) than one indicate that the robust estimator coefficient is larger (smaller) in magnitude compared to the standard method. In our preanalysis plan we selected 10% as our threshold for a substantively significant difference between the methods (i.e., $m = 0.90$ or $m = 1.10$).

4.1 Results

We find that the CVMF test selects the robust estimator as the better fit in six studies (38%). The 95% HPD interval generated from the Bayesian bootstrap for this estimate is [15%, 59%] and 86% of its posterior density exceeds our preregistered substantive threshold of 25%. Thus, there is some support for the claim that the robust estimator is relevant to political science (according to the CVMF test). However, we do not find that the robust estimator is the better choice at the same rate as DH; 97% of the posterior density falls *below* the value of 60% that they report.

Figure 1 presents the distributions of absolute coefficient ratios for the DH replication studies as well as our preregistered sample of studies. The graph illustrates the TCRA evaluation approach in our framework; it presents the replication context in the form of a summary of sample data. Most of the ratios are concentrated between zero and four, with some outliers extending to greater than 25. Thus, Figure 1 suggests that the robust estimator tends to produce coefficient estimates that are notably larger in magnitude compared to the standard estimator. Additionally, it shows

2 This illustration presents us with an important tension. We criticize the proof of concept standard in methodological replication analyses, but our own replication of a replication is itself just one example. We view it as an evaluation of our evaluation framework using the TCRA mode described above, albeit with a very small sample (one data point). While we cannot say much about the generalizability of our framework for replication analyses, we do claim that we avoided the selection bias problem by preregistering the example.

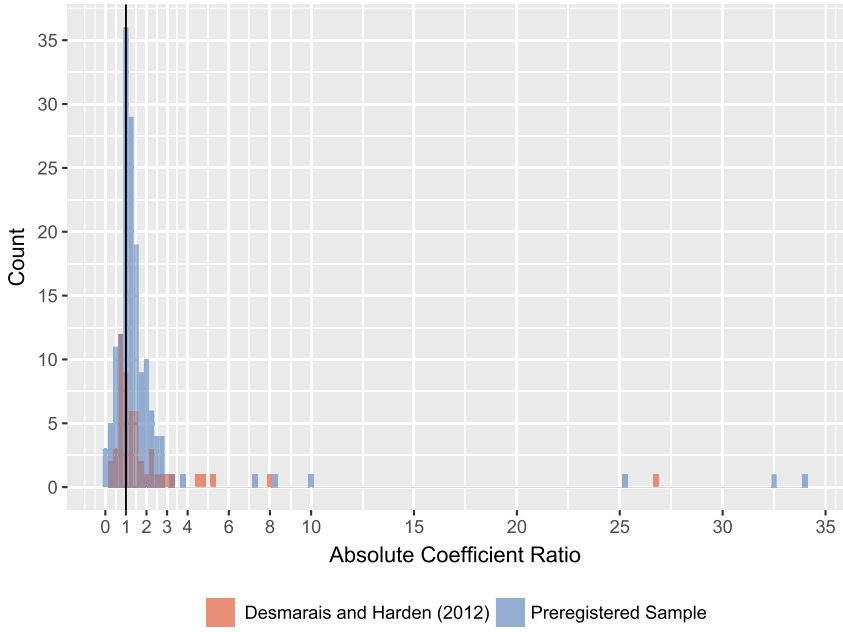

Figure 1. Distributions of absolute coefficient ratios. The graph presents the distributions of absolute coefficient ratios for the Desmarais and Harden (2012) replication studies and the preregistered sample of studies.

that the two distributions are fairly similar. The studies that DH chose to replicate are not obvious outliers on this metric.

We formally test the null hypothesis that the two methods are not different on average in Figure 2. That graph summarizes the posterior distributions for the mean absolute coefficient ratio in our full sample (all 209 coefficient estimates from all 16 studies), as well as separately for the DH sample and our preregistered sample. The points are posterior means and the line segments represent 95% HPD intervals. The results make a strong case for substantive differences between the two estimators. The average ratio for the full sample is 2.01, indicating that the robust estimator produces estimates that are, on average, double the magnitude of the standard estimator. Moreover, in the full sample and the two subsets *all* of the posterior density is larger than zero *and* larger than our substantive threshold of 1.1. These patterns indicate that meaningful substantive differences between the estimators are quite likely in a typical political science application. Finally, we again see that there is not a large difference between DH's sample and our preregistered sample.

In sum, our replication of replications confirms one of DH's findings, but shows less evidence for the other. Consistent with the original replication results, we find that coefficients from the two estimators are likely to yield notably different substantive conclusions. However, applying the CVMF test to a larger sample of studies suggests that the robust method is not the better-fitting estimator as frequently as DH imply. Their original replication analysis may *overstate* the robust estimator's practical value in this regard.

This example highlights the advantages of our evaluation framework. By assessing the methods with a representative sample of replication studies, we gain more insight into their practical utility. Furthermore, documenting and justifying our sample of studies in advance of conducting the replications improves the credibility of our comparisons. We did not start with the assumption that the robust estimator is substantively different from and/or better than the standard estimator, then seek only (or mostly) confirming evidence. Instead, we began with the null hypothesis that the two estimators are equal, then collected data with which to test that null.

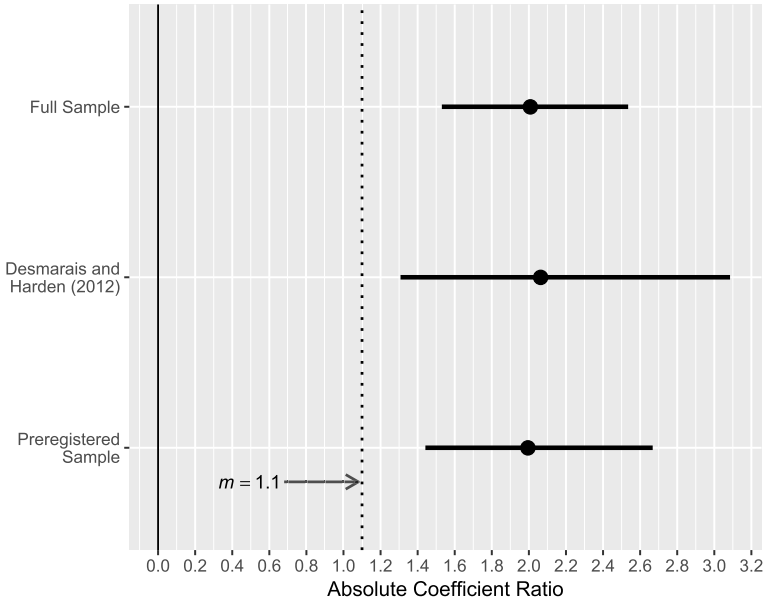

Figure 2. Posterior summaries of mean absolute coefficient ratios. The graph presents posterior means (points) and 95% HPD intervals (line segments) for the mean absolute coefficient ratio in each sample.

5 Conclusions

This letter describes an alternative framework for evaluating the practical utility of new methods. Instead of showing one or two replication studies as most-likely cases with a proof of concept evaluation standard, we advocate preregistering (many) replication studies, then describing and/or drawing inferences from that sample. Adopting the framework in practice will require more work from methodologists. However, we believe that the advantages of setting a stricter evaluation standard for new methods justify these added costs because they will further increase the valuable role that political methodology plays in the discipline. Moreover, we expect that these costs will decrease over time. As more journals establish and implement standardized practices for archiving replication data, the catalog of available replication studies will grow and large-scale replication analyses will become more feasible.

Thinking in terms of a sample of replications need not preclude the inclusion of a few illustrative examples, which we believe have pedagogical and substantive value. However, emphasizing a distribution of results from a sample is logical because it aligns with how researchers establish support for substantive claims. Furthermore, a shift in focus from one or two studies to many studies will naturally reduce anxiety about the "gotcha" dynamics that are becoming increasingly relevant in replication studies (Berinsky et al. 2018). In addition to improving the evaluation of new methods, our framework can promote replication as a collective benefit for the scientific community rather than a punishment for a few of its members.

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2018.54.

References

Bednarski, Tadeusz. 1993. Robust estimation in Cox's regression model. *Scandinavian Journal of Statistics* 20(3):213–225.

Berinsky, Adam, James N. Druckman, and Teppei Yamamoto. 2018. Why replications do not fix the reproducibility crisis: A model and evidence from a large-scale vignette experiment. Paper presented at the 5th Annual Asian Political Methodology Meeting, Seoul, South Korea.

Carsey, Thomas M. 2014. Making DA-RT a reality. *PS: Political Science & Politics* 47(1):72–77.
Cox, David R. 1975. Partial likelihood. *Biometrika* 62(2):269–276.
Desmarais, Bruce A., and Jeffrey J. Harden. 2012. Comparing partial likelihood and robust estimation methods for the Cox regression model. *Political Analysis* 20(1):113–135.
Gerring, John, and Lee Cojocaru. 2016. Selecting cases for intensive analysis: A diversity of goals and methods. *Sociological Methods & Research* 45(3):392–423.
Gill, Jeff. 1999. The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52(3):647–674.
Harden, Jeffrey J., Anand E. Sokhey, and Hannah Wilson. 2018. Replication data for: Replications in context: A framework for evaluating new methods in quantitative political science, https://doi.org/10.7910/DVN/ADJWFS, Harvard Dataverse, V1.
Ishiyama, John. 2014. Replication, research transparency, and journal publications: Individualism, community models, and the future of replication studies. *PS: Political Science & Politics* 47(1):78–83.
King, Gary. 1995. Replication, replication. *PS: Political Science & Politics* 28(3):444–452.
Monogan, James E. 2013. A case for registering studies of political outcomes: An application in the 2010 House elections. *Political Analysis* 21(1):21–37.
Rainey, Carlisle. 2014. Arguing for a negligible effect. *American Journal of Political Science* 58(4):1083–1091.
Rubin, Donald B. 1981. The Bayesian bootstrap. *The Annals of Statistics* 9(1):130–134.