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

Supplemental Appendix

Jeffrey J. Harden* Anand E. Sokhey† Hannah Wilson‡

September 3, 2018

Contents

1 Meta Analysis **1**
1.1 Coding . . . 1
1.2 Results . . . 2

2 The Evaluation Framework **3**
2.1 Preregistration . . . 5
2.1.1 Define the Population . . . 6
2.1.2 List the Target Sample . . . 7
2.1.3 Define the Quantities of Interest . . . 9
2.2 Conducting and Reporting the Replications . . . 11
2.2.1 Inference with Replication Results . . . 12

3 Cox Model Example **18**
3.1 Methodological Background . . . 18
3.1.1 Summaries of Methods . . . 19
3.2 Implementing the Evaluation Framework . . . 20
3.2.1 Defining the Population . . . 20
3.2.2 The Target Sample . . . 21
3.2.3 Replication Details . . . 22

*Associate Professor, Department of Political Science, University of Notre Dame, 2055 Jenkins Nanovic Halls, Notre Dame, IN 46556, jeff.harden@nd.edu.

†Associate Professor, Department of Political Science, University of Colorado Boulder, 333 UCB, Boulder, CO 80309, anand.sokhey@colorado.edu.

‡Graduate Student, Department of Political Science, University of Notre Dame, 2060 Jenkins Nanovic Halls, Notre Dame, IN 46556, hwilson2@nd.edu.

1 Meta Analysis

In the main text we summarize results of a meta analysis of methodological journal articles that employ replication analyses as part of their justification for a new method's value. We provide additional details on that meta analysis here. We collected articles published between January 2008 and June 2018 in all regular issues of *American Political Science Review*, *American Journal of Political Science*, and *Political Analysis*. Using the journals' respective websites, we first read the abstract and text of the article to determine whether it introduced a new method *and* employed a replication analysis. This process produced a total of 78 articles: 5 from *American Political Science Review*, 24 from *American Journal of Political Science*, and 49 from *Political Analysis*.

1.1 Coding

After identifying this sample of articles, we next moved to the coding process.[1] We first coded the number of replications performed in each article (or its appendix). Next, we coded whether the article included justification of the sample of replication studies in any capacity. Articles that included any discussion of case selection, however brief, were coded as having justified their selections. Next we moved to coding at the replication study level. Specifically, we coded the results of each replication study within each article using the following categories. Note that the first four categories are mutually exclusive, while the category "reversal" is a subset of "weaker" and "mixed."

- *Weaker* (65 replication studies). This category included any replication study in which all of the hypotheses received weaker support with the new method.
- *Stronger* (22 replication studies). This category included any replication study in which all of the hypotheses received stronger support with the new method.
- *Mixed* (39 replication studies). This category included any replication study in which some of the hypotheses received weaker support, some received stronger support, and/or some

[1] Coding instructions and a spreadsheet with our coding results is included in the replication materials for this letter.

received no change in support with the new method.

- *Same* (43 replication studies). This category included any replication study in which all of the hypotheses received no change in support with the new method.
- *Reversal* (23 replication studies). This category included any replication study in which one or more findings were fully reversed with the new method.

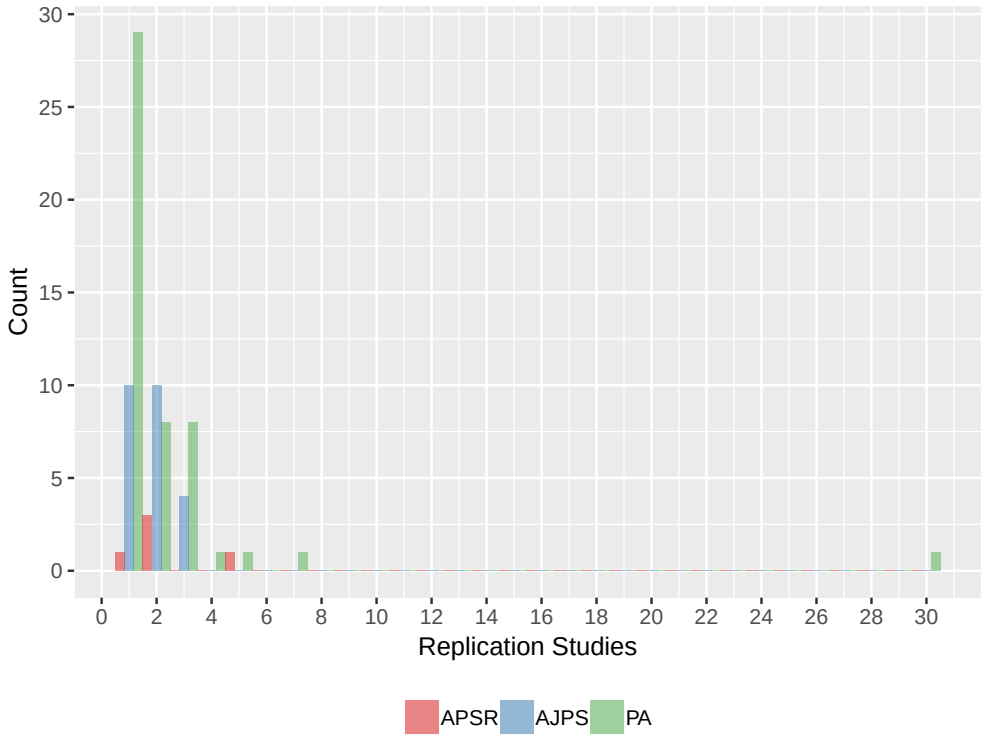
1.2 Results

As we note in the main text, the results of our meta analysis suggest that some of the problems with methodological replication analyses that we discuss are quite common. Consider selection bias, for example. Although methodologists may use replications to justify new methods in a variety of contexts, this meta analysis shows that they do so most often when the proposed method produces different results than those reported in the initial research.[2] Indeed, the vast majority of replications in our sample show some degree of divergence from the originally published results. Moreover, these articles seldom justify the cases selected for replication, even despite our generous coding standard in this regard. We are unable to observe how many replication studies the authors of each article considered as they conducted their research. However, the combination of a disproportionate amount of replication studies showing weaker or mixed results and a general lack of justification for the samples of replication studies is consistent with the selection bias problem.

We also note that generalizability is a concern in the current conduct of replication analyses. Figure A1 presents, for each journal, the distribution of replication studies reported in articles coded in our meta analysis. It shows that methodologists typically report a small number of studies. There are some exceptions to this pattern. Lall (2016), for instance, reports 30 replications (in a sample of studies that is clearly justified). But that example represents a clear outlier in comparison to the rest of the articles we coded. The mean is only 2.20 and the standard deviation is 3.38. The median, which is perhaps a better measure of location in this case, is just one replication study.

[2]This finding is consistent with the "gotcha bias" problem that plagues replication analyses in general (see Berinsky, Druckman, and Yamamoto 2018).

Figure A1: Distributions of Replication Studies Reported in Methodological Articles, 2008–2018



Note: The graph presents the distributions of replication studies reported in articles coded in our meta analysis.

Even if no selection bias is present, these small samples of replication studies limit what methodologists can say about a method's broader impact on the research community. Thus, we see considerable value in attempting to increase the number of replication studies in methodological articles, particularly given that the discipline as a whole is moving toward more widespread availability of replication data.

2 The Evaluation Framework

Here we detail the steps involved in our framework for evaluating new methods. Recall that the two modes of evaluation that comprise the framework share a common set of initial steps, but differ in the extent to which they facilitate generalizable claims about the utility of the new method. Importantly, both modes offer a more comprehensive evaluation of a new method compared to the proof of concept approach that is the current standard in methodological replication analyses. In

brief, the methodologist begins both modes by defining a relevant population of applied research studies and collecting a sample of replication data from that population. He or she documents this information in a preanalysis plan, then carries out the replications. The two modes diverge at this point; one involves description of the replication results while the other involves conducting inference about the new method's effect on applied research.[3]

The first mode of our framework, the test of concept replication analysis (TCRA), is similar to the proof of concept method in that the methodologist uses the replications as illustrative examples of the new method. The key difference between the current approach and TCRA stems from the fact that the replication plan is preregistered in the latter, which alleviates concerns about selection bias. The methodologist conducts the replications as described in a preanalysis plan, then presents a complete picture of the results, likely with histograms or summary statistics of key quantities. In other words, the goal of this mode is to provide description of a preregistered, larger, more representative sample of replication studies. It stops short of projecting from the in-sample replication results to the broader population of research studies. Thus, although it sets a higher evaluation standard than the proof of concept approach, there is still a natural ceiling on how much information the TCRA mode can convey to applied researchers. However, it may be all that is feasible if, for example, replication data are hard to obtain.

The second mode—the full inference replication analysis (FIRA)—includes all of the steps of the TCRA, then adds one crucial additional step: it treats empirical evaluation of a new technique as an inference problem, similar to the manner in which applied researchers test substantive hypotheses. The FIRA mode involves using results from the replications as data in testing the null hypothesis that there is no substantive difference, on average, between the existing and new methods of analysis. Importantly, we recommend establishing *a priori* the substantive criteria for testing

[3]Our framework is limited to cases in which the methodologist can make a useful comparison between different methods. In some cases a new method might not provide a "better" estimate of the same quantity, but instead yield new quantities altogether that allow the analyst to test new hypotheses.

this null. Compared to the TCRA, this mode of replication can provide more analytic leverage regarding the new method's utility to applied researchers. However, it is also more demanding with respect to the availability of replication data. In the remainder of this section we further elaborate on the details of both of these modes of replication.

2.1 Preregistration

The crucial first step in our evaluation framework—for both the TCRA and FIRA modes—is to publicly describe the replication plan before collecting or analyzing any data. Preregistration of research designs has gained popularity in political science in recent years (e.g., Humphreys, Sanchez de la Sierra, and van der Windt 2013; Monogan 2013). The practice promotes transparency and honesty by compelling scholars to commit to key elements of the research design before collecting and analyzing the empirical evidence. Doing so reduces the chances of a researcher making design choices that reflect "data fishing," or searching for results that are perceived to increase the likelihood of publication. Preregistration is not without its own problems (see Laitin 2013; Gelman 2013), but it does provide a useful means of reducing the adverse incentives that researchers face when analyzing data.

We contend that preregistering replication analyses is a vital component to the introduction of new methods. Preregistration directly addresses one of the main problems we describe in this letter: selection bias in the set of replication studies that methodologists present. Current practice opens up the possibility that a methodologist could privately conduct a large number of replications, but only publicly report one or two. This approach is, in our view, quite problematic; omitting many replication studies that are relevant to the new method could bias the research community's view of that method. For instance, reporting only those studies that show large differences between the existing and new methods could produce overconfidence in its utility to applied researchers.

In contrast, a preanalysis plan for replications establishes which studies are relevant in advance of looking at the data. This approach reflects an important change to the goal of the replication enterprise. Instead of focusing on finding one or two examples that demonstrate that the new method *can* have a big effect on substantive research (i.e., the proof of concept standard), it emphasizes

the need to place replications in context. Evaluating the impact of a new method on a collection of several replication studies chosen ahead of time provides applied researchers with more valuable information and sets a stricter criterion for the method's practical utility.[4]

The contents of the preanalysis plan will likely vary to some degree depending on the details of the method under study. However, we propose three key elements to include in any application of this approach. The first two of these include defining the population of replication studies relevant to the new method and describing the target sample of studies drawn from this population for actual replication. Additionally, we advise defining the specific quantities that will be computed as part of the replication process to measure the differences between the existing and new methods.

2.1.1 Define the Population

Placing replications in context involves clearly defining the relevant population of studies that are potentially affected by the new method. This population may vary in size, but will ideally be quite general. We recommend considering a broad definition of the types of studies that may have benefited from the new method if it had been available at the time they were being conducted. We encourage methodologists to consider multiple subfields of the discipline and perhaps even move beyond political science entirely.[5] Examples might include "all studies in political science that employ dyadic data" or "all studies that estimate the Cox proportional hazards model." Alternatively, the methodologist might decide to attempt a replication of all studies in a specific journal (Dewald, Thursby, and Anderson 1986) or a specific literature (Lall 2016).

This step is a significant one because it communicates the scope of the new method's potential impact on the research community. If the method has wide-ranging reach into numerous literatures

[4]Of course, a preanalysis plan alone does not guarantee a representative sample of replication studies.

[5]For simplicity we focus only on a population of political science studies in our example, but this distinction may not be necessary or advisable. Indeed, many new methods are broadly applicable across the social sciences and beyond. Thus, methodologists should also consider the impact of their innovations beyond political science.

and empirical strategies, defining this population serves as an initial piece of evidence in favor of the utility of the new method for applied researchers. That is not to say that methodologists should avoid developing methods for smaller, specialized research audiences. However, we contend that the scope of the problem is relevant information when evaluating methodological innovations. A large population of potential studies that might be affected represents a good preliminary sign of a new method's importance.

Methodologists should also keep the "file-drawer problem" (e.g., Franco, Malhotra, and Simonovits 2014) in mind when defining the population. The set of studies that ultimately makes it to publication is not random, which may affect generalizability. If possible, obtaining replication files for unpublished work would be ideal, but such data will likely not be available. Under such a case, the methodologist should note that the population of interest comprises *published* studies that employ the methods of interest.

2.1.2 List the Target Sample

Replicating all of the studies that fall into the population will likely never be feasible. Thus, another piece of information that methodologists should include when preregistering replication analyses is their plan for collecting a sample of replication studies from the population. Practical matters may limit these efforts to some extent. For example, if the population is defined over many disciplines, the methodologist might only be able to sample a few studies from each one. A recency bias is also likely to appear due to the fact that widespread availability of replication data is a relatively new phenomenon in political science. Even generating a list of all studies in the population is unlikely to be feasible, so drawing a simple random sample from the population is not a practical goal.

Nonetheless, it should still be possible to construct a sample that is representative of the population of studies. To do so, we first recommend searching a large database of academic articles (e.g., Google Scholar or JSTOR) for keywords that identify studies from the population. We advise refining this search to look within journals that have publicly-available replication archives in the timeframe for which replication data are available. For instance, searching for articles published

since 2014 in journals that have adopted the Data Access & Research Transparency (DA-RT) initiative will facilitate easy acquisition of data once the methodologist is ready to begin conducting the replication analysis.[6] These choices will likely weight the search results more heavily toward recent research. However, we expect that this bias will naturally decline over time as more journals require replication data to be deposited prior to publication.

The keywords employed in the search should be carefully selected so that they accurately distinguish studies that employ the existing method of interest. However, some false positives that result from overly general search terms are preferable to specific terms that miss relevant studies. The methodologist should gather the entire list of results returned by the search—which we refer to as the sampling frame—then read through each article to verify that it belongs to the population of interest. We refer to this subset of relevant articles as the target sample. Note that this verification should be done using only the text of each article; it should *not* involve the replication data or software code. We recommend reporting bibliographic information of all of the studies in the sampling frame and clearly identifying which studies comprise the target sample.

In an ideal scenario the target sample can be credibly deemed representative of the population of relevant studies. If representativeness is in doubt, the methodologist should note that and discuss the relevant issues. For instance, there may be overrepresentation of recent work or work from specific subfields. Variation in the implementation of replication policies by journals may also create problems for sampling. Regardless of the source of the problem, a key consideration when one arises is whether it is correlated with the specific quantities the methodologist intends to collect. For instance, are studies from recent years more likely to show large differences between the two methods compared to studies from many years ago? If the methodologist has reason to suspect that such a pattern may appear, he or she should document the issue as well as any empirical strategy

[6]In political science, a number of journals that represent much of the empirical work in the discipline signed the Journal Editors' Transparency Statement (JETS) in 2014 as part of DA-RT. By signing JETS, an editor publicly declared several commitments in support of a transparent research process, including easy access to replication data as a requirement for publication.

for correcting it.

If replicating every study in it is not feasible, the list could be further narrowed by selecting a random subset, so long as replication code for the random selection is provided in the preregistration document. Most importantly, documenting the sampling frame and the target sample ahead of time helps alleviate selection bias in the replication process. These lists establish all of the replication studies the methodologist *could have conducted*, given the keywords used in the search. Documenting this information sharply contrasts with the current approach, in which methodologists report a small number of studies from an unknown sampling frame.

Finally, we recommend listing which models or statistical analyses within a given study the methodologist plans to replicate, whether it is all of them or a subset.[7] Doing so addresses the fact that the selection of models for replication is itself a researcher degrees of freedom problem. If the methodologist chooses a subset of models to replicate, he or she should describe the selection criteria at this point as well.[8]

2.1.3 Define the Quantities of Interest

Another piece of information that a preregistration of a replication analysis should contain is a list of quantitative measures that the methodologist will use to compare the original results with

[7]If feasible, replicating all of the models in each study may be the optimal choice. Doing so removes a key researcher degree of freedom; the methodologist does not have to make choices of which models to replicate and which ones to ignore. Additionally, it provides more information for evaluating the methods, which increases the precision of the methodologist's conclusions. However, the main drawback to this approach is that it could lead to unequal weighting of replication studies, because some articles report many specifications and some report few. To avoid this problem, the methodologist may want to choose one "main" specification from each study (and preregister his or her criteria).

[8]In our preanalysis plan for the example described here we preregistered the criteria for replication model selection, but not the actual list of models, which we added post-hoc (see Table A2). We thank an anonymous reviewer for suggesting this step.

those from the new method. We refer to these measures as replication quantities of interest (RQI). Defining the RQI is important because it clearly communicates the specific criteria by which the new method will be judged. In other words, it tells the research community how it will know if the impact of the new method is "small" or "large." The methodologist should include as much detail as possible on the measures and justification for why they are useful in evaluating the influence of the new method on substantive conclusions.[9]

The RQI will depend on the application. If the new method is an alternative estimator of a statistical model, one option might be the ratio of a coefficient estimate from the new method to the same estimate from the existing method (see our example in the main text). This RQI gives the proportional change in the estimate between the two methods. It is useful because it is easily comparable across many replication studies and can be applied to several parameter estimates or a specific subset.[10] Additionally, a coefficient ratio has substantive appeal for applied researchers whose primary concern is explanatory and/or causal modeling. In essence, it efficiently answers the question "how different would substantive conclusions about the effect of this covariate be if the new method was used instead of the existing method?" Changes in the statistical significance of key parameter estimates between existing and new methods may also be relevant (e.g., the ratio or difference in p-values). However, as a general rule we advise the use of RQI that address substantive significance.[11] Finally, methodologists may be interested in comparing some type of

[9]We also recommend establishing which statistical models or analyses within a given study the methodologist plans to replicate.

[10]We do not recommend using a linear difference between two estimates unless the estimates fall on a scale that is common across replication studies.

[11]Another potentially informative RQI could be an indicator of whether an estimate from the new method is statistically significantly different from the analogous estimate with the existing method. This quantity would be most relevant if the methodologist views the new method as *complementing* the existing method; that is, his or her advice to applied researchers would be to report results from both methods. If the new method is meant to entirely *replace* the existing method—perhaps because the existing method has some fundamental flaw—this RQI may not be

predictive loss function between the two methods (e.g., correct classification metrics). These quantities could be used with the framework as well. Our main advice is simply that methodologists should declare and justify the RQI ahead of time.

The central objective of the FIRA mode in our evaluation framework is to conduct a test of the null hypothesis of no average substantive difference between the existing and new methods using the RQI. This can be done with null hypothesis significance testing (NHST), in which case we advise declaring a significance threshold in the preanalysis plan. However, NHST can be quite restrictive (see Gill 1999), so we also advise that methodologists pay attention to substantive differences. The process we recommend for doing so, which we detail in section 2.2.1 below, involves selecting a cutoff value that denotes the smallest substantively meaningful RQI value. This value can and should be informed by substantive context, but will always be somewhat arbitrary. Thus, we advise selecting it ahead of time and reporting it in the preanalysis plan as well.

2.2 Conducting and Reporting the Replications

Once the preanalysis plan is deposited, the next step is to obtain the corresponding replication files from each study in the target sample (or the random subset) and verify that the results can be replicated using those files. We recommend that the research paper include the list of studies that clear this final check for replicability. The accountability mechanism of preregistration is strongest if there is transparency regarding the sampling frame and target sample (see above) as well as the list of studies that ultimately appear in the data. After verifying replicability, the methodologist can employ the new method. He or she should seek to replicate the original results exactly and ensure that the new method is conducted on the exact same data as the existing method. We advise reporting and explaining any deviations from this norm. After completing this step, the chosen RQI should be computed according to the preanalysis plan. Depending on the research goals, there may be several RQI generated from a single replication dataset. This process should be repeated for every replication study in the final list, and the RQI stored in a common data file. Once this file is complete, the final step in the process is the evaluation and reporting of differences between

as useful.

existing and new methods.

If the methodologist chooses the TCRA mode, comparing the methods under study consists of providing descriptive information about the RQI. Graphical summary methods such as histograms and boxplots might be useful, as well as numerical quantities (e.g., means, medians, and standard deviations). The objective in this case is to emphasize the replication context rather than just one or two examples. Importantly, both the location and scale of the distributions of RQI are relevant for the comparison; providing them further elevates the evaluation standard of this mode above a proof of concept. Consider a proof of concept replication analysis that only reports replications with large differences between existing and new methods. In such a case the methodologist only informs applied researchers about a tail of the RQI distribution, and applied researchers may not even know that they are only seeing extreme values. In contrast, the TCRA approach provides a complete picture of the distribution of replication results. However, the TCRA approach is still illustrative in nature. We next turn to the FIRA mode, which allows methodologists to make more generalizable claims about the impact of a new method on applied research.

2.2.1 Inference with Replication Results

The FIRA mode employs the logic of hypothesis testing to evaluate new methods from a sample of replication results. The null hypothesis that we consider is that there is no difference, on average, between the existing and new methods. For example, without loss of generality consider a RQI discussed above: the ratio of two methods' coefficient estimates, with each covariate indexed by i, for J replication studies indexed by j.

$$r_{ij} = \frac{\widehat{\beta^{ij}_{\text{new}}}}{\widehat{\beta^{ij}_{\text{old}}}}. \tag{1}$$

Testing for a statistical difference between the existing and new methods based on this RQI implies the following null (H_0) and alternative (H_A) hypotheses regarding the mean of r, which we denote

τ:

$$H_0\colon\ \tau = 1 \tag{2}$$
$$H_A\colon\ \tau \neq 1.$$

Testing this null requires a theory of inference. The usual framework of frequentist hypothesis testing (e.g., t-tests) could be used in this scenario, but would be problematic for two reasons. First, the standard version of this approach requires a normality assumption, which may be suspect if the sample size is not large enough to invoke the central limit theorem (CLT). This concern is particularly relevant here given that replication data can be hard to obtain, and thus the sample size may be small. Consider a typical "rule of thumb" for the necessary sample size to justify the CLT: 30 observations (e.g., Urdan 2005). Reaching that number is often feasible in data collection for a substantive study, but collecting enough replication studies to yield 30 values of the RQI may be more difficult, especially if the chosen RQI is defined at the study level.

Second, a standard frequentist hypothesis test forces the methodologist to rely on test statistics, p-values, and the language of statistical significance to measure the new method's value. We contend that this approach is too restrictive in general (see Gill 1999), and especially so for the assessment of new methods. A p-value can only communicate the likelihood that the data are consistent with the null hypothesis; it cannot tell the analyst much about the credibility of an alternative hypothesis. Yet, the credibility of an alternative is exactly the information that would benefit applied researchers the most in this context. Specifically, researchers need to know something about the relative value of the new method compared to the existing method. We advocate communicating this information with nuance and detail rather than a simple statement about whether or not the difference between the existing and new techniques is statistically significant.

We address both of these issues—the suspect nature of the normality assumption and the need for more information in interpretation—by bootstrapping from a Bayesian perspective, with the goal of constructing a posterior distribution for τ (e.g., Rubin 1981; Efron 2012). For a vector of data values x of length n, classical bootstrapping can be conceptualized as repeatedly drawing

probability weights $\pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$ for the n data points and computing the statistic of interest (e.g., the mean of x) weighted by π. The weights are constructed by drawing counts of each observation in a bootstrap sample from a multinomial distribution with n trials and all probabilities $p \in (p_1, p_2, \ldots, p_n)$ equal to $\frac{1}{n}$, then normalizing such that the counts sum to 1.[12] The Bayesian bootstrap accomplishes this same objective, but draws the weights from a uniform posterior Dirichlet distribution with parameter $\alpha = 1$ (see Rubin 1981).[13]

Classical and Bayesian bootstrapping are essentially equivalent in practice, especially as n grows larger.[14] Indeed, classical bootstrapping is akin to generating a "poor man's Bayes posterior" (Hastie, Tibshirani, and Friedman 2009, 272). However, the Bayesian approach is useful here for two reasons. First, in small samples—which are likely in this context—the classical bootstrapping weights generated from normalizing discrete multinomial counts can be "choppy" due to the presence of very few observed values of x. The Dirichlet distribution, by contrast, allows for smooth draws of the weights at all sample sizes. Second, bootstrapping in a Bayesian framework

[12]For example, consider a vector of four data points. One possible draw of counts from the multinomial distribution would be (1, 0, 2, 1), which would translate to weights of (0.25, 0.00, 0.50, 0.25). In other words, that sample would give no weight to observation #2 (i.e., leave it out of the sample) and double the weight of observation #3 compared to unity (i.e., observation #3 appears twice in the sample).

[13]The sampling distribution for π remains a Multinomial$(p_1, p_2, \ldots, p_n)$ as in the classical case. The posterior is the product of this sampling distribution and a limiting Dirichlet prior, which concentrates density near the vertices of an n-dimensional simplex as $\alpha \to 0$ (Rubin 1981). This prior is weakly informative in that it gives the most weight (but not all) to x values that appear in the observed data. In fact, classical bootstrapping is a special case of the Bayesian bootstrap with a Dirichlet prior and $\alpha = 0$, which places all of the prior weight on the observed x values. The posterior of the statistic of interest is generated by repeatedly sampling π from its posterior and using each draw as a weight in the computation.

[14]Bayesian bootstrapping as we describe here is straightforward in the R statistical environment with the `bayesboot` package (Bååth 2016; R Core Team 2018).

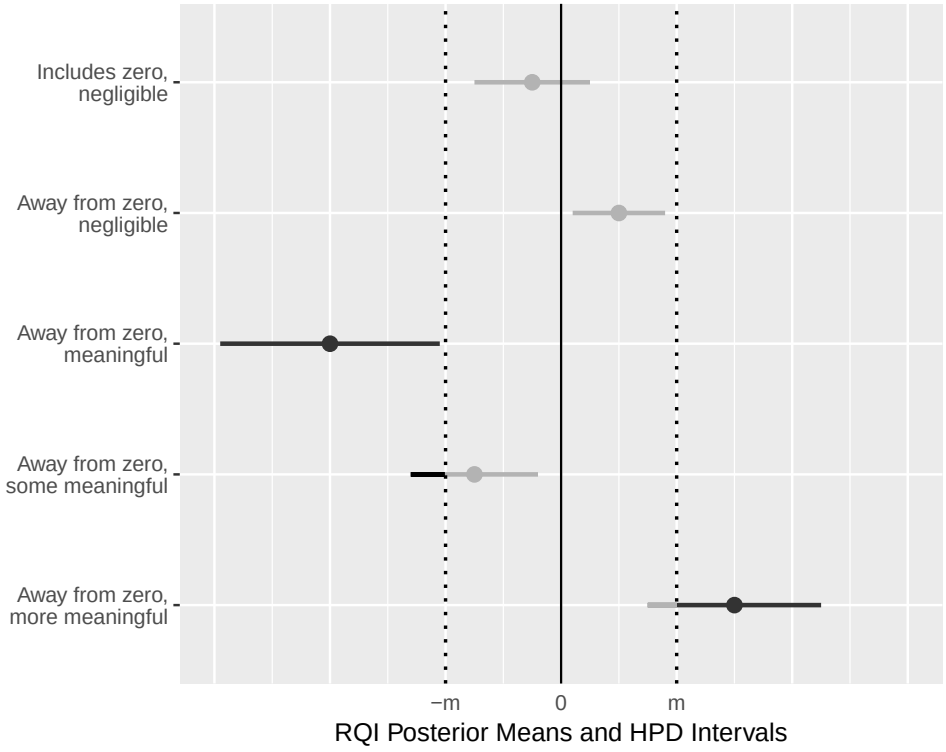
produces a proper posterior distribution of the statistic of interest to evaluate in hypothesis testing. Instead of simply reporting whether the mean RQI is statistically significant or not, the methodologist can employ the more substantively intuitive logic of Bayesian inference to draw conclusions about the new method.

We advise formally assessing substantive differences between the existing and new methods via Bayesian highest posterior density (HPD) intervals generated from the posterior distribution of τ. To do so, we recommend Rainey's (2014) method of evaluating whether "meaningful effects" are conceivable in a given sample, modified for the Bayesian context.[15] This approach can be used to determine if "large" differences—as defined by the RQI—between the two methods' results are plausible. A key aspect of this process involves choosing a cutoff value, which Rainey (2014) refers to as m, that defines the smallest substantively meaningful average RQI. In other words, the methodologist should determine how different, on average, the existing and new methods must be such that it would be inadvisable for an applied researcher to ignore the new method as a possible empirical strategy. This quantity can be symmetric such that values greater than m and less than $-m$ are deemed meaningful, but symmetry is not required (Rainey 2014, 1085). The choice for m is a somewhat arbitrary one (but see Rainey 2014, 1085), which is why we advise selecting it ahead of time and documenting the choice in the preanalysis plan.

With m in place, the next step is to compare the HPD interval for τ to m. The critical question to answer is how much of the HPD interval is contained inside $[-m, m]$. As more posterior probability falls inside that range, meaningful differences between existing and new methods are less likely, on average. Alternatively, if large portions of the posterior exceed m (or fall below $-m$), relatively more plausible values of τ represent substantively meaningful differences. Figure A2 presents several illustrative examples. The points plot hypothetical posterior means for a given RQI mean and the lines indicate 95% HPD intervals. Vertical dotted lines represent $-m$ and m and the solid

[15] Rainey's (2014) focus is on how applied researchers can use frequentist confidence intervals to argue in favor of null effects; here we adapt his approach to use HPD intervals in assessing substantively meaningful differences.

Figure A2: Illustrative Examples of Mean RQI Posterior Interpretations



Note: The graph presents example scenarios for interpreting the posterior distribution of the mean RQI. The points plot hypothetical posterior means for a given RQI mean and the lines indicate Bayesian HPD intervals. Vertical dotted lines represent $-m$ and m and the solid line indicates zero (i.e., no difference between the two methods). Portions of each HPD interval that fall inside (outside) $[-m, m]$ are colored gray (black).

line indicates zero (i.e., no difference between the two methods). Portions of each HPD interval that fall inside (outside) $[-m, m]$ are colored gray (black).

Interpretation of the first case is straightforward; it shows a HPD interval that is entirely in the range of $[-m, m]$ and overlaps zero, providing clear evidence of no meaningful difference between existing and new methods. The next example is what would be referred to as statistically significant from a frequentist perspective. However, it is not substantively significant under our application of Rainey's (2014) method. The entire HPD interval is bounded away from zero but still falls entirely inside $[-m, m]$. In contrast, the third example shows a statistically *and* substantively significant τ; the HPD interval is entirely to the left of $-m$. That case provides clear evidence of a notable difference between existing and new methods.

The final two examples highlight the value of employing Bayesian inference. From a frequentist perspective, the two cases would be considered essentially identical: both estimates are statistically significant (HPD intervals bounded away from zero) and neither estimate is substantively significant (HPD intervals inside $[-m, m]$). However, the sizes of the portions of the posterior distributions that fall outside of $[-m, m]$ indicate a noteworthy distinction between the two cases. The posterior distribution of the former case includes approximately 30% of its density to the left of $-m$, while about 86% of the latter distribution is larger than m. Thus, while both HPD intervals are bounded away from zero, the last example shows a great deal more evidence that the average RQI is substantively large.

This FIRA mode represents our ideal standard for evaluating new methods. Reporting these inferences—along with the descriptive measures of the TCRA mode—yields a great deal of useful information for applied researchers. Most importantly, it provides an authentic picture of the new method's expected value to the research community. The FIRA mode is much more demanding than the proof of concept approach in that it requires the completion of several steps ahead of time as well as execution of a comprehensive set of replications. However, in our view this additional work is worth the resulting increase to the standard for substantive evaluation of new methods.

A final important point to note is that our framework could be extended beyond the relatively simple comparison of RQI that we describe above. Specifically, instead of just collecting the RQI, the methodologist could also collect additional information from each study as "replication covariates." Then those variables could be used in the analysis of the RQI. Using them to weight studies based on some relevant criteria is one option. Another might be to move to a regression framework. That is, regress the RQI on the other features of the studies, such as sample size, correlation structure between variables, or number of covariates in the model. This approach would give the methodologist insight into what empirical conditions might lead us to expect large or small differences between the methods, which would allow him or her to provide very precise advice about their substantive consequences.

3 Cox Model Example

This section contains additional details about our example of the evaluation framework: a replication of the replication analysis in Desmarais and Harden (2012). We provide more information about the estimators in the comparison, discuss our implementation of the evaluation framework, and supplement our previous summary of the results. Much of this content can also be found in our preanalysis plan.

3.1 Methodological Background

The Cox proportional hazards model is a popular choice among several alternatives for the analysis of duration data, likely due to its inherent flexibility (Box-Steffensmeier and Jones 2004). However, Desmarais and Harden (2012) note that the method is particularly sensitive to deviations from its assumptions, such as measurement error or omitted variable bias. These problems can easily generate outlying event times (observations with large deviations between the actual rank of the duration and the model's expected failure rank), which bias parameter estimates. Bednarski (1993) presents an alternative estimator of the Cox model that downweights outliers with the goal of reducing this bias. Desmarais and Harden (2012) compare this alternative estimator to the conventional partial likelihood estimation method (Cox 1975) and develop a sample-based test—which they call the cross-validated median fit (CVMF) test—for empirically choosing between the two approaches.

In our replication of replications, we compare Bednarski's (1993) estimator and the conventional partial likelihood approach. Desmarais and Harden (2012) conduct this same comparison, reporting results from five replication studies. We employ our evaluation approach to examine the practical utility of the alternative estimation method based on a larger, more representative sample of replication studies. We examine the full distribution of differences between the two methods and conduct inference regarding the null hypothesis that the alternative estimator is no different, on average, from the conventional approach. Additionally, we assess the frequency with which Desmarais and Harden's (2012) CVMF test selects the alternative estimator. The results place the

replications of Desmarais and Harden (2012) in better context, which is especially important here because they focus their attention on the replications that point to the alternative method as the better choice.

3.1.1 Summaries of Methods

The standard partial likelihood method, or PLM, is the original approach to Cox model estimation (Cox 1975). It involves maximizing a "partial" likelihood function that is described as such because only the ranks of the durations are used, not the actual event times. This estimator converges to maximum likelihood as the sample size increases if the model's assumptions are valid (see Desmarais and Harden 2012, 115). However, if one or more of the Cox model's assumptions are *not* valid, the PLM approach will yield sensitivity in its estimates. The iteratively reweighted robust (IRR) estimator is designed to address this problem (Bednarski 1993). Specifically, it modifies the score of the partial likelihood to downweight outliers, thereby reducing their influence on the final estimates. Outliers in this context are defined as observations with large differences between the expected and actual failure ranks; they often result from misspecification problems such as measurement error or omitted variables (Desmarais and Harden 2012, 115–116). The analyst controls the severity of the downweighting by setting a proportion (e.g., 0.05) of the most extreme outliers to obtain weights of zero.

The PLM and IRR estimators represent a classic bias-variance tradeoff. PLM is the more efficient estimator of the two, but exhibits greater bias under misspecification. The IRR estimator reduces bias, but at the expense of efficiency. The CVMF test, which is the main innovation proposed in Desmarais and Harden (2012), is designed to help researchers navigate this tradeoff in the context of a sample of data. Specifically, it provides information about which estimator—PLM or IRR—fits the observed data better. In brief, the test evaluates the two estimators by comparing their observation-wise contributions to the log-partial likelihood. It computes these contributions via leave-one-out cross-validation. Then, it constructs a binomial test statistic as a count of the observations for which the IRR cross-validated log-partial likelihood is greater than that of PLM. The null hypothesis of equal fit implies a binomial distribution of size equal to the estimation

sample size with probability of success equal to 0.50. Thus, the researcher can obtain a p-value to test the null of equal fit.[16]

3.2 Implementing the Evaluation Framework

In our evaluation framework the first step is to preregister the replication plan. In our preanalysis plan we (a) defined the relevant population of studies, (b) listed the target sample from this population, and (c) defined RQI.[17] We discuss each of these items below.

3.2.1 Defining the Population

The relevant population for this replication analysis is any *published* political science study that employs the Cox model.[18] This is a broad definition, covering all empirical subfields of the discipline. For simplicity, we limit our scope to political science, but our definition could cover other fields as well. We do not include studies that *only* employ parametric duration models. While the data in these studies could be used to estimate the Cox model, we exclude them because those researchers chose a different methodological strategy. Studies that employ parametric duration models *and* the Cox model are included in the population.

[16]For example, consider a hypothetical sample of 100 observations. The null hypothesis of equal fit stipulates that the expected number of IRR log-partial likelihood contributions greater than PLM log-partial likelihood contributions is 50. If the data show that 60 observations obtain larger IRR log-partial contributions, the test statistic is 60 and the corresponding p-value based on a $\mathscr{B}(100, 0.50)$ distribution is ≈ 0.02. Using a standard significance threshold of 0.05, this outcome would result in a rejection of the null hypothesis of equal fit and the selection of the IRR estimator as the better fit for that particular sample and model specification.

[17]We deposited the preanalysis plan at the Political Science Registered Studies Dataverse on June 22, 2018 (https://doi.org/10.7910/DVN/J7HFRX).

[18]In our preanalysis plan we declared the population to be any political science study that employs the Cox model. However, we did not sample from unpublished studies when collecting data, and so we must limit this definition to published work only.

3.2.2 The Target Sample

The large size of the population of studies defined above necessitates that we draw a sample of studies with which to conduct our replication analysis. This sample should be representative of the population, so ideally we would randomly sample from the entire population. However, a simple random sample is not feasible because we do not have a list of every study that belongs in the population and replication data are not available for all studies. Accordingly, we first selected a set of journals, then randomly sampled studies from those journals. We had two goals in mind when choosing journals: covering a representative collection of empirical political science and selecting journals for which we knew replication data are publicly available.

Specifically, we began with the list of 27 journals that signed the Journal Editors' Transparency Statement (JETS) as part of the discipline-wide Data Access & Research Transparency (DA-RT) initiative in 2014.[19] This choice ensured that we would be able to find studies with available replication data. To obtain representative coverage of the empirical subfields of political science, we chose the four most prominent general interest journals from the JETS list, which also happen to be among the most prominent journals in the discipline: *American Political Science Review* (APSR), *American Journal of Political Science* (AJPS), *Journal of Politics* (JOP), and *British Journal of Political Science* (BJPS). We also selected three subfield journals: *State Politics & Policy Quarterly* (SPPQ, American politics), *Comparative Political Studies* (CPS, comparative politics), and *Journal of Peace Research* (JPR, international relations).

Next we accessed each journal's public replication data repository and determined the time period for which data were available for all articles published that year (as of June 22, 2018). We report this information in the first two columns of Table A1. To generate the sampling frame, we first searched each journal in Table A1 in the years listed using Google Scholar. We used the search string [Cox proportional hazards model].[20] The results—comprised of a total of 33 studies—are

[19] See the statement here: https://www.dartstatement.org/2014-journal-editors-statement-jets.

[20] We also tried "Cox model" alone and with "Cox proportional hazards model". The search string we chose appeared to be the most effective at minimizing false positives and eliminating

listed in a spreadsheet file in the replication materials. Next we read each article and denoted in the file whether it is relevant (i.e., belongs in the population). To make this determination we examined each article's text for positive evidence of the use of the Cox model for at least one model mentioned in the article.[21] For instance, word searches for "Cox" typically pointed us to phrases in the text or tables of results presenting Cox model estimates. This process yielded our target sample: 24 relevant studies that we could potentially replicate, as described in Table A1.

Table A1: Time Periods of Full Data Availability in Journal Replication Data Repositories

Journal	Website	Years	Search Hits	Relevant Studies
APSR	https://dataverse.harvard.edu/dataverse/the_review	2017–2018	1	1
AJPS	https://dataverse.harvard.edu/dataverse/ajps	2013–2018	7	4
JOP	https://dataverse.harvard.edu/dataverse/jop	2015–2018	5	4
BJPS	https://dataverse.harvard.edu/dataverse/BJPolS	2015–2018	5	2
SPPQ	https://dataverse.unc.edu/dataverse/sppq	2015–2018	2	1
CPS	http://journals.sagepub.com/home/cps	2014–2018	2	2
JPR	https://www.prio.org/Data/Replication-Data/	2000–2018	11	10

Note: Cell entries report replication data information for each journal that we sampled. Search hits comprise the sampling frame; relevant studies comprise the target sample. The search result information is current as of June 22, 2018.

3.2.3 Replication Details

We attempted to replicate the entire target sample and re-replicate all five replication studies presented in Desmarais and Harden (2012). Our objectives with each study were to (1) reproduce the PLM Cox model results from a "main" model specification reported in each article, (2) estimate

false negatives.

[21]The model itself could be reported in the article or an appendix, but we required that it be a reported model instead of a vague appeal to robustness.

the same model with IRR, and (3) conduct the CVMF test.[22] To start this process we read each article and determined which model was of central substantive interest to the original authors; we wanted to avoid underspecified models, such as those that only include the main covariate of interest without control variables. For articles presenting several regression models (e.g., with and without control variables), we identified the main regression by determining which one the authors refer to the most in their discussion of results. In some cases this identification was straightforward (some articles only report one Cox model) while in others it was more challenging. If more than one model appeared to be discussed approximately the same amount, we selected the first one presented in the relevant table in the article. Table A2 reports which model we chose for each replication study and the rationale behind each choice.

We were ultimately successful in completely replicating a total of 16 studies: 11 from the target sample and the five Desmarais and Harden (2012) replication studies.[23] Table A2 lists the entire set of studies and information about our replication efforts, including results from the CVMF test. We include brief notes explaining the results of each study. Those labeled "exact replication" produce exact matches between the results reported in the original articles and our own estimation. In some cases we found slight differences, likely due to software discrepancies. Other times, practical matters forced us to alter the specification slightly (e.g., remove a covariate or fixed effects). Finally, in some cases the original model was not capable of estimation with IRR, such as frailty models or models with interval-censored or counting process data (denoted "outcome variable incompatible with IRR").

We focus on two RQI in our evaluation of the methods. First, we record the CVMF test's

[22]Throughout this replication analysis we set the IRR truncation parameter—which defines the proportion of observations with weights of zero—to the software default of 0.05. We accepted any other choices (e.g., the procedure for handling tied durations) that the original studies report.

[23]We successfully replicated the original PLM version of the model in all but three of the articles. However, in some of these we were unable to complete IRR estimation and/or the CVMF test (see Table A2).

Table A2: Replication Studies and Outcomes

Article	Journal	Model	Rationale	Sample	Successful Replication	CVMF (p-value)	Notes
Box-Steffensmeier et al. (1997)	APSR	Table 2 (331)	Selected by Desmarais and Harden (2012)	DH	✓	IRR (0.12)	Exact replication
Martin (2004)	AJPS	Table 1, Model 2 (454)	Selected by Desmarais and Harden (2012)	DH	✓	PLM (0.00)	Exact replication
Mattes and Savun (2010)	AJPS	Table 2 (520)	Selected by Desmarais and Harden (2012)	DH	✓	IRR (0.09)	Exact replication
Martin and Vanberg (2003)	BJPS	Table 2, Model 3 (330)	Selected by Desmarais and Harden (2012)	DH	✓	IRR (0.00)	Exact replication
Golder (2010)	CPS	Table 2, Model 4 (20)	Selected by Desmarais and Harden (2012)	DH	✓	PLM (0.00)	Exact replication
Graham et al. (2017)	APSR	Table A7 (Appendix)	Only Cox model reported	Preregistered	✓	IRR (0.00)	Must exclude covariate `bmr2_prevauth` from the specification for IRR to estimate successfully
Park and Hendry (2015)	AJPS	–	–	Preregistered	–		Methodological article
Kelley and Simmons (2015)	AJPS	Table 1, Model 1.3 (64)	First of several models most often discussed in the article	Preregistered	–		No replication materials
Laver and Benoit (2015)	AJPS	Table 5, Model 3, Column 5 (286)	Model most often discussed in the article	Preregistered	✓	PLM (0.74)	Exact replication
Thrower (2017)	AJPS	Table 1, Model 3 (651)	First of several models most often discussed in the article	Preregistered	–		Sample size too large for IRR to converge
Camerlo and Prez-Lin (2015)	JOP	Table 1, Model 3 (615)	Model most often discussed in article	Preregistered	–		Replication materials do not produce reported results
Davis and Wilf (2017)	JOP	Table 1, Model 3 (973)	Model most often discussed in the article	Preregistered	–		Outcome variable incompatible with IRR
Johns and Pelc (2018)	JOP	Table 1, Column 3 (11)	First of two models most often discussed in the article	Preregistered	✓	IRR (0.00)	Must exclude covariates `hhi_trade_t`, `startyear`, and legal issue fixed effects from the specification for IRR to estimate successfully
Boudreau and MacKenzie (2018)	JOP	Table A3, Column 1 (Appendix)	Only Cox model reported	Preregistered	✓	PLM (0.00)	Exact replication
von Stein (2016)	BJPS	Table 1, Model 1 (666)	First of three models most often discussed in the article	Preregistered	–		Outcome variable incompatible with IRR
Ecker and Meyer (2017)	BJPS	Table S3, Model 6 (Appendix)	Only Cox model reported	Preregistered	✓	IRR (0.00)	Must exclude year fixed effects from the specification for IRR to estimate successfully
Cayton (2016)	SPPQ	Table 2, Model 2A (81)	First of several models most often discussed in the article	Preregistered	–		No replication materials
Crespo-Tenorio et al. (2014)	CPS	Table 2, Model 1 (1061)	First of several models most often discussed in the article	Preregistered	–		Frailty model not available with IRR
Turcu and Urbatsch (2015)	CPS	Table 1, Model 5 (420)	Model most often discussed in article	Preregistered	–		No replication materials
Krustev (2006)	JPR	Table 1, Model 5 (254)	Model most often discussed in the article	Preregistered	✓	PLM (0.00)	Slight differences in replicated and reported coefficient estimates
Meernik and Brown (2007)	JPR	Table 2 (76)	Only Cox model reported	Preregistered	✓	PLM (0.89)	Slight differences in replicated and reported coefficient estimates
Quackenbush and Venteicher (2008)	JPR	Table 3, Model 8 (734)	Model most often discussed in the article	Preregistered	–		Outcome variable incompatible with IRR
Nilsson (2008)	JPR	Table 1, Model 1 (488)	First of several models most often discussed in the article	Preregistered	–		Outcome variable incompatible with IRR
Cunningham (2010)	JPR	Table 1, Model 4 (123)	First of several models most often discussed in the article	Preregistered	–		Outcome variable incompatible with IRR
Kim (2012)	JPR	Table 2, Model 3 (313)	First of two models most often discussed in the article	Preregistered	–		Outcome variable incompatible with IRR
Acosta (2016)	JPR	Table 3, Model 1 (189)	First of three models most often discussed in the article	Preregistered	✓	IRR (0.00)	Exact replication
Lektzian and Regan (2016)	JPR	Table S2 (Appendix)	Only Cox model reported	Preregistered	✓	PLM (0.10)	Slight differences in replicated and reported coefficient estimates
Tokdemir and Akcinaroglu (2016)	JPR	Table 4, Model 1 (275)	First of two models most often discussed in the article	Preregistered	✓	IRR (0.00)	Slight differences in replicated and reported coefficient estimates
Rudloff and Findley (2016)	JPR	Table 2, Model 4 (26)	Model most often discussed in the article	Preregistered	✓	IRR (0.73)	Exact replication

Note: Cell entries report information about replication efforts with all potential studies in the replication analysis. CVMF test results reflect two-tailed p-values.

selection (TS): the estimator selected by the test for the model of interest. This quantity is measured at the article level (sample size of 16) with a categorical variable that takes on the values PLM, IRR, or neither (i.e., equal fit). We identified m = 25% in our preanalysis plan as the threshold for a substantively meaningful proportion of studies for which IRR is the better estimator.[24] We employ two-tailed tests of the null hypothesis of equal fit between the two estimators with the conventional (and preregistered) $\alpha = 0.05$ threshold. Desmarais and Harden (2012) use a more relaxed standard. In their claim that the test selects IRR in three of the five studies, they report p-values below 0.05 for two replications (Martin and Vanberg 2003; Mattes and Savun 2010) and p = 0.06 for a third one (Box-Steffensmeier et al. 1997). However, best practice would likely dictate the use of two-tailed tests in all cases unless there was strong theoretical or empirical reasons to prefer one estimator *a priori*. We summarize the results with 10,000 draws from the posterior of each RQI mean (i.e., 10,000 bootstrap replicates).

As we discuss in the main text, the CVMF test selects the robust estimator as the better fit in six studies, or 38% (95% HPD interval: [15%, 59%]). This value exceeds the threshold of 25% that we established in our preanalysis plan as substantively meaningful, although it is much smaller than the 60% that Desmarais and Harden (2012) report (97% of the posterior density falls below 60%). If we switch to Desmarais and Harden's (2012) criteria—one-tailed tests and $\alpha = 0.06$—the CVMF test selects IRR in 8 studies, or 50% of the sample (95% HPD interval: [27%, 74%]), which is closer to (but still below) their reported value (78% of the posterior density falls below 60%).

The second RQI we compute is the ratio of the absolute values of the IRR coefficient estimates to the absolute PLM estimates. We refer to this quantity as the coefficient ratio (CR). It is equal to 1 if the IRR and PLM estimates are equivalent, greater than 1 if the IRR estimate is larger in magnitude, and less than 1 if the PLM coefficient is larger in magnitude. This RQI provides a

[24] We decided that it would not be necessary for the robust estimator to be the better option a majority of the time, as Desmarais and Harden (2012) seem to imply. Instead, we wanted to choose a value that would reflect a notable minority of studies. The specific value of 25% captures this criterion, although it also reflects some anchoring toward an intuitive number.

measure of the magnitude of the difference between the two estimation methods. We chose m = 10% as the threshold for a meaningful difference between estimators.[25] It is defined at the article-coefficient level, yielding a sample of 209 observations. The results reported in the main text and discussed below reflect bootstrapping at this level. However, we also conducted the bootstrapping at the article level to account for the possibility of nonindependence between coefficients from the same article. Results are substantively similar (see the replication materials).

On average, the results with this quantity suggest that the IRR estimation method could produce substantively distinct conclusions. As we note in the main text, the average ratio for the full sample is about 2.01 (95% HPD interval: [1.53, 2.54]), or 101% larger magnitude of coefficients from the robust estimator. The posterior density completely exceeds our preregistered substantive threshold of 1.1. Furthermore, the HPD intervals indicate that average ratios of up to about 3 are plausible. It is also important to note that these results are similar between the Desmarais and Harden (2012) sample of replication studies and our new preregistered sample. While Desmarais and Harden (2012) may overstate the value of their CVMF test (see above), their replications and our new set clearly show that these two estimators can produce meaningful substantive differences.

[25]This choice was rooted in our own past experiences as applied researchers dealing with model dependence. We decided that even a difference as small as 10% in the magnitude of two estimates can be substantively consequential.

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 File Drawer Problem: A Model and Evidence from a Large-Scale Vignette Experiment." Presented at the 5th Annual Asian Political Methodology Meeting, Seoul, South Korea.

Box-Steffensmeier, Janet M., and Bradford S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. New York: Cambridge University Press.

Bååth, Rasmus. 2016. "bayesboot: An Implementation of Rubin's (1981) Bayesian Bootstrap." R package, version 0.2.1. https://CRAN.R-project.org/package=bayesboot.

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.

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

Efron, Bradley. 2012. "Bayesian Inference and the Parametric Bootstrap." *The Annals of Applied Statistics* 6(4): 1971–1997.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345(6203): 1502–1505.

Gelman, Andrew. 2013. "Preregistration of Studies and Mock Reports." *Political Analysis* 21(1): 40–41.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3): 647–674.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learn-*

ing: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21(1): 1–20.

Laitin, David D. 2013. "Fisheries Management." *Political Analysis* 21(1): 42–47.

Lall, Ranjit. 2016. "How Multiple Imputation Makes a Difference." *Political Analysis* 24(4): 414–433.

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.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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.

Urdan, Timothy C. 2005. *Statistics in Plain English*. 2nd ed. London: Lawrence Erlbaum Associates.