Beyond the Hazard Ratio: Generating Expected Durations from the Cox Proportional Hazards Model

Jonathan Kropko* Jeffrey J. Harden†

August 6, 2015

Abstract

The Cox proportional hazards model is commonly used in duration analyses. However, because it is estimated using only the observed durations' rank ordering, typical quantities of interest (QI) come from the hazard function, such as hazard ratios or percentage changes in the hazard rate. These QI are easy to misinterpret, substantively vague, and difficult for many audiences of social science research to understand. We propose the COX ED method, which estimates expected durations from the Cox model by fitting a generalized additive model (GAM) of the observed durations on the ranks of the linear predictor values. This allows researchers to calculate expected durations and marginal changes in duration for a specified change in a covariate. These QI closely match researchers' theoretical interests and are easily understood by most readers. We validate COX ED through a simulation study, then employ it in re-analyses of published articles from three subfields of political science.

Keywords: Cox proportional hazards model; Survival models; Generalized additive models; Quantities of interest; Expected durations

*Assistant Professor, Department of Politics, University of Virginia, S383 Gibson Hall, 1540 Jefferson Park Avenue, Charlottesville, VA 22904, jkropko@virginia.edu.

†Assistant Professor, Department of Political Science, University of Colorado Boulder, 416 Fleming, UCB 333, Boulder, CO 80309, jeffrey.harden@colorado.edu.

1 Introduction

The Cox proportional hazards model is a popular method of duration analysis that has been employed in every empirical subfield of political science. For example, the Cox model has been used to study the time required for coalition government formation in multiparty democracies (Diermeier and van Roozendaal 1998; Martin and Vanberg 2003), delay in the U.S. Senate's confirmation of federal judges (Binder and Maltzman 2002; Shipan and Shannon 2003), challenger entry into U.S. House races (Box-Steffensmeier 1996), position-taking on legislation in Congress (Box-Steffensmeier, Arnold, and Zorn 1997), the duration of militarized conflicts (Krustev 2006; Meernik and Brown 2007), peace after wars (Fortna 2004; Mattes and Savun 2010), and many other political processes. However, in spite of its well-earned popularity, the standard method of reporting results from the Cox model is easy to misinterpret, substantively vague, and difficult for many audiences of social science research to understand. In this paper we detail these problems and provide a solution.

Duration models (also called survival models), are built around the concept of hazard, which represents the risk that an event will occur (e.g., "failure") at a particular point in time given that it has not occurred (or failed) up to that point. There are two widely-used, general classes of duration models that make different statements about hazard. One class, the class of parametric duration models, begins with an assumption about the general shape of the baseline hazard for every observation over time. For instance, the exponential model makes the assumption that the baseline hazard is constant, the Weibull model assumes it increases or decreases monotonically, and the log-normal model assumes it is either monotonic or increases towards a single mode and decreases thereafter (Box-Steffensmeier and Jones 2004).

The second class, the class of semi-parametric duration models, includes the Cox proportional hazards model (Cox 1972, 1975). The Cox model does not make an assumption about the shape of the baseline hazard, which gives it considerable flexibility. For this reason, the Cox model has become a preferred option for researchers in several fields of study. To avoid an assumption about the baseline hazard, the Cox model disregards the magnitudes of the event times and instead only

considers their relative ranks, or the *ordering* of the cases based on their observed durations. The use of ranks alone allows the Cox model to maximize a partial likelihood function to estimate coefficients without having to include the baseline hazard function in the computation. In effect, the problem of characterizing the baseline hazard function is circumvented, not solved. As a result, estimates of expected duration or the marginal change in expected duration with respect to a change in a covariate are not readily available from the Cox model.

Instead, researchers typically make substantive interpretations of Cox model results via relative changes in the hazard function. For example, the coefficient estimates can be used to construct quantities of interest (QI) called hazard ratios that report the average multiplicative change in the ratio of each observation's hazard—denoted $h_i(t)$, where i is an observation and t is time—to the baseline hazard, $h_0(t)$, as a result of a one-unit increase in a covariate.[1] Applied researchers usually report hazard ratios with an emphasis on whether they are greater or less than one to describe the direction of an effect, then look to the p-value to test the null hypothesis that the ratio is equal to one (i.e., no effect).

QI from the hazard rate are mathematically correct, so we do *not* claim that researchers who employ them are necessarily making incorrect inferences. However, we contend that understanding the substantive implications of Cox model results could be greatly improved by shifting to QI based on expected durations, or the expected length of time until event occurrence, according to the estimated model. Applied research in political science appears to support this contention; below we show evidence from over 60 published articles employing the Cox model that researchers' hypotheses are more often focused on the time to event occurrence, not on the risk of event occurrence.

Furthermore, compared to expected durations, hazard rate QI present several complications with respect to substantive interpretation, which make communication of results to social scientists and (especially) general audiences difficult. Hazard ratios exist on the same scale as the probability

[1]These quantities are often referred to as "hazard ratios" when they actually represent the multiplicative change in this ratio. In that way, hazard ratios are similar to odds ratios from a logistic regression. Henceforth we follow previous work in referring to them as hazard ratios.

density function of failure over time for each observation, and therefore describe changes in the height of densities, not probabilities. Yet it is easy to mistakenly conceptualize them as changes in probability. Additionally, because QI based on the hazard function have no meaningful scale, questions regarding the magnitude of an effect are difficult to answer. Finally, even if used correctly, hazard ratios still require technical knowledge to understand, and therefore do not work well in presenting research to general audiences such as students, journalists, and policymakers.[2]

Our objective in this research is to derive a method for computing QI from the Cox model that are more intuitive, easier to interpret in terms of both the direction and magnitude of an effect, and straightforward for a general audience to comprehend. To that end, below we develop and validate a method for computing expected durations and marginal changes in expected duration, with estimates of uncertainty, from the Cox model. We call this method *Cox Proportional Hazards with Expected Durations*, or COX ED. This method is *not* a new estimator of the parameters of the Cox model (its first step is estimation of the Cox model just as researchers have always done). Rather, COX ED is a new approach for drawing substantively-meaningful inferences from Cox model estimates.

We motivate the need for the COX ED method in the next two sections. We use data collected from the text of articles in top journals to demonstrate that political scientists tend to frame their hypotheses in terms of duration, but then switch to discussing the risk of event occurrence after Cox model estimation. Then in section 3 we discuss several other shortcomings of QI based on the hazard rate. We discuss why past solutions to these problems are not ideal in section 4, then describe implementation of COX ED in section 5. In section 6 we evaluate COX ED in a simulation study and demonstrate its superior performance compared to parametric models (from which expected durations are readily available). We then apply COX ED in section 7 to replicate and extend four of the published studies listed above (Box-Steffensmeier 1996; Binder and Maltzman 2002; Martin and Vanberg 2003; Mattes and Savun 2010). We provide answers to substantively important questions that the Cox model cannot answer with hazard ratios alone:

[2]As King, Tomz, and Wittenberg (2000) point out, statistical models must "convey numerically precise estimates of the quantities of greatest substantive interest... and require little specialized knowledge to understand" (347).

- How many more days will it take for a government to form if the bargaining parties are ideologically distant?
- How much longer will the Senate delay the confirmation of a judge if government is divided?
- By how many weeks can an incumbent delay a quality challenger's entry into a race if she raises more campaign funds?
- How much longer will peace last after a civil war as a result of an uncertainty-reducing provision in the peace agreement?

Finally, in section 8 we discuss practical issues with implementing COX ED in applied work and conclude.

2 How Do Researchers Use the Cox Model?

Before describing the COX ED method in detail, we establish the need for a method of generating expected durations from the Cox model.[3] We accomplish this with a systematic assessment of how researchers employ the Cox model in substantive work. Specifically, we conducted a meta analysis of journal articles that report one or more Cox models appearing between 1990 and 2015 in four leading political science journals: *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, and *International Organization*. We addressed two main questions in this analysis: (1) What kind of language do researchers who employ the Cox model tend to use in framing their hypotheses? (2) What method(s) do these researchers use to communicate results of the Cox model?

2.1 Framing Hypotheses

We began by identifying all articles in our chosen time period that reported the estimation of the Cox model in the main text.[4] This produced 63 articles across the four journals ranging in publication date from 1996–2015. Next, we collected the text of the articles' hypotheses. We then identified words in this text as part of either a "risk frame," a "duration frame," or not related to framing. We dropped most of the words because they fell into the third category (i.e., they were

[3] We assume reader familiarity with duration models. See the appendix for a brief summary of these models or Box-Steffensmeier and Jones (2004) for a more comprehensive treatment.

[4] See the appendix for full details of this analysis.

not relevant to the manner in which the authors' framed their hypotheses). For the words that did fit a category, we coded any word that related to probability, likelihood, or chance in the risk frame and any word relating to time in the duration frame category (see the appendix for the complete list of words).

For example, consider Maltzman and Shipan's (2008) analysis of the stability of legislation in Congress, which contains a representative example of a hypothesis framed in terms of risk. The authors posit that "Laws enacted during periods of divided government are more *likely* to be amended than those enacted under unified control" (256, emphasis added). The key term in this hypothesis is *likely*, signaling that the authors are interested in how covariates (such as divided government) influence the likelihood, or risk, that the phenomenon of interest (whether a law is changed) occurs. In contrast, Lo, Hashimoto, and Reiter (2008) report their expectations using the duration frame. They hypothesize that "Peace following interstate war *lasts longer* when the war ends with one state suffering foreign-imposed regime change" (720, emphasis added). Here the authors frame their expectation in terms of how *long* the phenomenon of interest (the duration of peace) will last as a function of a covariate (foreign-imposed regime change).

In all, we coded eight unique words as risk frame words and 47 unique words as duration frame words. The counts of total words strongly favored the duration frame (215 words) over the risk frame (119). We also counted the unique and total words from each frame within each article, then coded the article as either predominantly using a risk frame, duration frame, or equal use of both frames. Using the unique word count, we coded 33 articles as a duration frame, 16 with a risk frame, and 14 with equal use of both frames. With the total word count these numbers were 33, 20, and 10, respectively.

Thus, the first part of our meta analysis revealed that political scientists employing the Cox model over the last 25 years tend to discuss their theoretical expectations in the language of time until event occurrence. Language related to the risk of event occurrence also appears, but it is less common than duration-based framing. Approximately 70% of the hypotheses in articles in our sample contain more duration words or an equal amount of duration and risk words. In contrast,

only 48% contain more risk words or an equal number from the two frames. It is clear that researchers' substantive interests usually center on the duration of political phenomena, not just their likelihood of occurring.

2.2 Interpretation Methods

Our second objective was to code which method(s) each article in our sample used to interpret the results of the estimated Cox model. This was accomplished by reading the results section and identifying each unique method employed. We created a total of four categories based on what we found in the text, which we list below with their frequencies (see the appendix for specific coding details).[5]

- Hazard ratios (24 articles).
- Changes to the hazard rate (33 articles).
- Empirical estimates of the hazard and/or survivor functions (17 articles).
- Only sign and significance of the coefficient estimates (8 articles).

All of the articles in our sample that go beyond sign and significance in their interpretation of the Cox model focus on the hazard rate, whether through hazard ratios, changes to the hazard rate, or estimation and graphing of the baseline hazard function. The closest that any article comes to interpreting results with respect to duration is the few (3 articles) that construct estimates of the survivor function. However, even in those cases the focus is on the probability of survival, not estimates of expected duration.[6]

To summarize, while researchers' hypotheses are often framed with respect to time, no article published in top political science journals in the last 25 years generates expected durations from the Cox model. Researchers who employ the Cox model are typically forced to switch the manner in which they discuss their research when moving from hypotheses to results. This motivates our

[5]The articles employed an average of 1.3 of these interpretation methods. 45 articles used one method, 17 articles used two methods, and 1 article employed three different methods.

[6]Berliner and Erlich (2015) report the "half-life" of an event for a given configuration of covariates, which is defined as the time point in which the expected probability of event occurrence reaches 0.50. This is potentially useful for interpreting substantive effects, but is not exactly the same as estimating an expected duration.

research, which provides a method for generating expected durations from the Cox model. The COX ED approach described below allows researchers to maintain consistency between the language they use to describe their theoretical framework and the language they use to communicate their empirical findings.

3 The Hazards of Hazard Ratios

As our meta analysis of journal articles shows, QI from the hazard rate are the accepted standard method for interpreting the results of a Cox model, despite the fact that most researchers are actually interested in the duration of political phenomena. Furthermore, even if researchers do not wish to frame their hypotheses with respect to time, we contend that hazard rate QI are limited in several other ways. First, we show that a single hazard ratio can correspond to many different ratios of failure *probability*. Second, we make the case that a hazard ratio or change in the hazard rate is a substantively vague quantity to which even experts may struggle to give precise meaning. Finally, we argue that even if a researcher can adequately explain the substantive implications of a hazard ratio, many important consumers of political science research may still have difficulty understanding it. In contrast, an expected duration is a straightforward, intuitive quantity that better communicates substantive significance to a wide range of readers.

3.1 The Relationship Between Risk and Failure Probability

Box-Steffensmeier and Jones (2004, 15) describe the substantive meaning of a hazard rate as "the risk a unit incurs of having a spell or duration end in some period, given that the spell or duration has lasted up to or beyond some length of time." The word risk conveys a positive relationship between hazard and failure probability, but the concept of risk itself is vague. This may lead researchers to mistakenly interpret results in probabilistic terms. If risk and probability were always the same, it may not be as necessary to look for other quantities to compute from the Cox model because probability is an intuitive concept that conveys precise substantive meaning and is easy for many readers to understand. However, although the two concepts are directly related, in general risk is *not* equal to failure probability.

In the appendix we show proof that a single hazard ratio almost never corresponds to a single probability ratio. In fact, a given hazard ratio usually corresponds to many probability ratios, depending on the duration under consideration and on the functional form of the baseline probability CDF. Therefore, the information conveyed by the hazard ratio essentially never gives the change in probability of event occurrence. If the baseline hazard function is known, the multiplicative change in the conditional probability of instantaneous failure could be calculated for any duration. However, for the Cox model there is no assumed parametric baseline hazard function, so we cannot know how the hazard ratio relates to multiplicative changes in probability. Researchers who employ the Cox model are limited to strictly reporting the effects of covariates on the risk—not probability—of event occurrence.

3.2 Intuition and Communication of Results

Another potential problem with QI generated from the hazard rate is that they can be substantively vague. Consider Shipan and Shannon's (2003) analysis of the duration of U.S. Supreme Court nominee confirmations. The authors report that, consistent with their expectations, when the opposing party of the president controls the Senate, the hazard rate of confirmation drops by 47.8% compared to when the president's party controls the Senate (665). While the authors' implementation and interpretation of the Cox model is methodologically sound, we contend that it is still difficult to put into precise substantive terms what a 47.8% reduction of hazard actually means. Is that estimate "large enough," or would it need to be greater than 50% (or 60%, or 75%, etc...) to be considered an "important" effect? Without a meaningful scale, that question is difficult to answer. The result of this substantive vagueness is that researchers can often only responsibly interpret the sign and significance of coefficient estimates and/or hazard rate changes.

Finally, even if a researcher is able to appropriately contextualize a hazard rate-based QI, the audience for which that discussion will make sense is primarily limited to other academics or those who have had graduate-level statistics training. This leaves out a wide range of potentially important consumers of the research findings. Students, journalists, and even policymakers may stand to benefit from the substantive conclusions that researchers make, but many lack the training

to comprehend statistical jargon and academic prose.

In contrast, an expected duration is a substantively intuitive concept that researchers can expand upon to add more nuance and detail. Shipan and Shannon (2003, 656) make the following substantive argument about the confirmation process:

> ...delaying the confirmation of a nominee may have important policy implications. Supporters of a nominee want to get him or her on the Court as soon as possible, so he or she can start influencing arguments, deliberations, and case decisions. Opponents, on the other hand, want to delay this process.

Expected durations would allow researchers to address this point directly. For instance, they could use Supreme Court caseload data to estimate how many cases a nominee would miss due to a confirmation delay. That allows for more specificity and depth in assessing model results, and puts researchers' substantive expertise to work (see our second replication below for an example). In addition, an expected duration is direct, intuitive, and requires virtually no specialized knowledge. We replicated Shipan and Shannon's (2003) analysis with our COX ED method described below. Nearly anyone can understand the substantive implications of our results: all else being equal, during divided government the confirmation of a nominee to the Supreme Court takes approximately 25 days longer than when the government is unified, give or take about 8 days.[7] Expected durations focus interpretation of the statistical model on the reason why scholars and non-experts alike care about the research: to understand the factors that affect the duration of important political phenomena.

4 Are There Existing Solutions?

It is important to note that we are not the first to point out the difficulties that arise with the interpretation and communication of Cox model results. Given the Cox model's heavy use in epidemiology and biostatistics, it is not surprising to find that researchers in those fields have also written on this issue (e.g., Bender, Augustin, and Blettner 2005). Hernan (2010) explains that the

[7] Of course, scholars could still disagree over whether 25 days represents a notable effect. But at least in that case the discussion would center on substantively meaningful quantities.

hazard ratio cannot be used for causal inference in medical studies, primarily because the hazard ratio may change over the lifespan of a patient (i.e., the proportional hazards assumption may be violated). On the issue of generating expected durations, he recommends avoiding the Cox model altogether and using a parametric model instead (14, see also Cox et al. [2007]).

Uno et al. (2014) also contend that hazard ratios are problematic because the proportional hazards assumption may be violated and because they cannot be translated "into a more transparent clinical benefit, such as the prolonged survival time" (2380). As an alternative, they provide several "model-free" means of analyzing survival data. These alternatives all involve comparisons of the survivor functions (estimated with Kaplan-Meier curves) of a treatment and control group. For example, they suggest computing the ratio of the survivor functions at a given point in time or the ratio of the median survival times in each group. These quantities are potentially useful in some contexts, but are generally most applicable to data generated from the experimental trials common in health sciences research.

Another possible solution that may be better suited for social science research is to use estimates of the baseline hazard or survivor function to generate expected durations (see Cox and Oakes 1984; Kalbfleisch and Prentice 2002; Collett 2003). While computationally feasible, this approach is not commonly employed in political science; as our meta analysis shows only a few articles report estimates of these functions at all.[8] Furthermore, this approach has, in our view, several suboptimal properties. It assumes that the baseline hazard is a step function, meaning there is a uniform probability of event occurrence between successive timepoints. It also considers the event occurrence rates in the sample to be estimates of the rates in the population, but does not account for uncertainty due to sampling variability. As we discuss below, COX ED does not assume uniform failure probability between observed failure times and does account for uncertainty due to sampling variability.

[8]To our knowledge, only Katz and Sala (1996) do something similar, and they report failure probabilities after estimating the baseline hazard, not expected durations.

5 Cox Proportional Hazards with Expected Durations

The goal of COX ED is to generate expected durations for individual observations and marginal changes in expected duration given a change in a covariate from the Cox model. Specifically, our method can compute the expected duration for each observation used to fit the Cox model, given the covariates, the expected duration for a "new" observation with an independent variable profile set by the analyst, or the first difference, or change, in expected duration given two new observations. Our software implementation of COX ED is a package called `coxed` in the R statistical environment.[9] The method proceeds according to five steps, which we detail below.

Step 1: Estimate a Cox model. The COX ED method uses coefficient estimates from the Cox model. Thus, researchers should first estimate the model just as they always have, paying close attention to measurement of variables, model specification choices, and other considerations. Issues such as the method for handling tied durations, testing the proportional hazards assumption, and model fit—while important on their own—should be resolved before implementing COX ED.

Step 2: Generate and rank the linear predictor. In this step the method computes expected values of risk for each observation by matrix-multiplying the covariates, X, by the estimated coefficients from step 1, β, then exponentiating the result. This creates $\exp(X\beta)$, or the linear predictor. Then the observations are ranked from smallest to largest according to their values of the linear predictor. This ranking is interpreted as the expected order of failure; the larger the value of $\exp(X\beta)$, the sooner the model expects that observation to fail, relative to the other observations.

Step 3: Fit a GAM. The next step is to connect the model's expected risk for each observation ($\exp[X\beta]$) to duration time (the observed durations). A generalized additive model (GAM) fits a model to data by using a series of locally-estimated polynomial splines (Beck and Jackman 1998). It is a flexible means of allowing for the possibility of nonlinear relationships between variables. COX ED uses a GAM to model the observed durations as a function of the linear predictor ranks generated in step 2. More specifically, the method utilizes a cubic regression spline

[9]Note that COX ED is different from the tools available in the popular `Zelig` and `simPH` packages in R (Imai, King, and Lau 2012; Gandrud 2015). The QI that these packages compute include the hazard ratio, survivor function, and hazard function. They do *not* compute expected durations from the Cox model.

to draw a "smoothed" line summarizing the bivariate relationship between the observed durations and the ranks.[10] A GAM is appropriate because the relationship between the observed durations and the ranks can be linear or nonlinear. It is similar to LOWESS methods (locally weighted scatterplot smoothing). The critical difference is that GAMs can generate expected values for new observations.[11]

Figure 1 shows an example of this step. The data and model come from Martin and Vanberg's (2003) research on the duration of coalition government bargaining, which is the first of our replication studies below. The graph gives the expected ranks of the observations on the *x*-axis—from the smallest values of the linear predictor on the left side of the graph (last to event occurrence) to the largest on the right (first to event occurrence)—against the observed durations on the *y*-axis.[12] The solid line represents the GAM fit and the shading indicates its 95% confidence interval. Note the clear downward (but nonlinear) relationship between the durations and the ranks. As an observation's value of the linear predictor becomes relatively larger (i.e., larger relative risk of event occurrence), its actual number of bargaining days decreases, but at a decreasing rate. This nonlinearity is captured by the GAM.

[Insert Figure 1 here]

Step 4: Generate expected durations and estimate marginal effects from the GAM fit. Expected durations can be computed for observations in the data, similar to generating expected values of the dependent variable from a linear regression model. To do this, COX ED uses the GAM fit to compute the expected value of the duration given the observation's rank. The solid line in Figure 1 shows these values for the Martin and Vanberg (2003) model. As an example, consider the observation highlighted in black (Netherlands, 1986). That observation's rank is 162 (*x*-axis),

[10] Several other smoothers are available in the R package `mgcv`, although we found minimal differences between them (for more details, see Wood 2006, 2011). The number of knots in the GAM is a tunable parameter in our R package.

[11] COX ED only uses observations that are not censored in step 3. Unlike the Cox model, the GAM has no means of accounting for censoring. Thus, using the observed durations of the censored observations could skew the fit of the GAM because those observations' durations are governed by the linear predictor *and* the limits of the data collection enterprise.

[12] No observations are censored in this example, but see our simulations and other replications for examples with censoring.

which corresponds to an expected duration of about 44 days according to the GAM (y-axis). Note also that the actual duration for that observation is 54 days (the black point). Thus, the GAM is off by about 10 days. As we discuss in section 6, we utilize the differences between the GAM's expected values and the actual durations as a means of assessing the performance of COX ED.

In order to examine marginal changes in duration given a change in a covariate, it is necessary to create two or more "new" observations corresponding to theoretically-interesting, hypothetical covariate profiles. For example, we might set an indicator variable to 0 and 1 or a continuous variable to a "low" and a "high" value.[13] For the other variables in the model, COX ED employs the "observed value" method of Hanmer and Kalkan (2013). Instead of setting those variables to their means or modes, it allows them to vary naturally over the entire data, then averages over them in the computations.[14] For instance, to estimate the effect of an increase in a covariate X_1 from 0 to 1 on the expected duration, we use the following steps:

(4a) Set X_1 to 1 for the entire data (all N observations) and calculate the linear predictor for every observation, then take the average value of those computations (the median is the default).

(4b) Repeat step (4a) while setting X_1 equal to 0.

(4c) Take the values obtained in steps (4a) and (4b) and append them to the list of linear predictor values from the original Cox model in which X_1 is left as exogenous data. Then compute new rankings of the linear predictor values from this list, which is length $N+2$.

(4d) Pass the list of rankings from step (4c) to the GAM from step 3 as new data to generate expected values. Note that a new GAM is not estimated at this step. Rather, expected durations are generated for each observation—including the two new ones created in steps (4a) and (4b)—using the previously estimated GAM. This produces point estimates of the expected durations for those two new observations.

[13] COX ED can also compute interactive effects by setting the constituent terms and the interaction term to desired values. For instance, consider the interaction effect with two indicator variables, X_1 and X_2. The proper interaction specification would include a parameter on each variable plus a parameter on the multiplicative term: $\beta_1 X_1 + \beta_2 X_2 + \beta_3(X_1 \times X_2)$. To compute the expected duration when $X_1 = 1$ and $X_2 = 1$, the analyst can easily set X_1 to 1, X_2 to 1, and $X_1 \times X_2$ to 1 in the `cox.ed()` function in our R package.

[14] This default can be changed at the discretion of the analyst.

(4e) Compute the difference between the two estimates obtained in step (4d): the expected duration for the data in which X_1 is set to 1 and the expected duration for the data in which X_1 is set to 0. This quantity is a point estimate for the marginal effect, or first difference, corresponding to the change in X_1 from 0 to 1.

Step 5: Repeat the process many times. To produce estimates of uncertainty, COX ED repeats steps 1–4 many times (1,000 is the default) via bootstrapping. The method generates bootstrap samples of the data and re-estimates the Cox model coefficients on each bootstrap sample.[15] At each iteration, this produces a new vector of actual durations and a new ranking of linear predictor values, which are then used to fit a new GAM. This results in a distribution of expected durations for each independent variable profile (e.g., step 4d) and a distribution of the marginal effect (step 4e). These distributions can be used to produce standard errors and confidence intervals for the estimates.[16] Importantly, by bootstrapping the entire process, this step incorporates the uncertainty from the Cox model estimation *and* the uncertainty from the GAM.

This process is mostly automated in our R package; analysts need only a `coxph` model object, the name of the variable of interest, and the two values of that variable they wish to input. However, the function also allows for several changes to default settings, including the formulation of the GAM and the computation of confidence intervals. Additionally, COX ED can also be used with models that include time-varying covariates (see our replication of Box-Steffensmeier [1996] in the appendix).

6 Monte Carlo Simulations

Having described the COX ED procedure, we next turn to an assessment of its performance. We evaluate COX ED based on its ability to return accurate expected durations and marginal changes in duration compared to three popular parametric models using simulated data. Importantly, we

[15] Standard bootstrapping at the observation level or cluster-level bootstrapping (see Harden 2011) are both available.

[16] By default, the method computes the standard errors of each quantity as the standard deviation of its bootstrap distribution. The halfwidth of the confidence interval is then computed by multiplying a tunable critical value based on the normal distribution by the standard error. The default critical value is 1.96 (i.e., a 95% confidence interval). A fully non-parametric confidence interval based on quantiles of the bootstrap distribution is also available.

simulate the data so as not to privilege the parametric models. We use a "random spline" data generating process (DGP), which generates baseline hazards by fitting cubic splines to randomly-drawn points. This produces a variety of shapes, some of which are monotonic or unimodal, but many of which are multimodal. Figure 2 displays four randomly generated baseline hazard functions using this method. See the appendix for details on this process as well as simulations with the assumed DGPs of the parametric models.

[Insert Figure 2 here]

We use these baseline hazard functions, along with three covariates and four true coefficients generated from standard normal distributions, to create simulated durations and marginal changes in durations that depend on data. We then run COX ED and the exponential, Weibull, and log-normal survival models on the simulated data, and assess how accurately they return the expected durations and marginal changes in duration.[17] We compare competing models because it is difficult to assess absolute performance in a simulation setting (Carsey and Harden 2013). We conduct these comparisons with two performance criteria: (1) the root mean square error (RMSE) of each method's expected durations for each observation and (2) the RMSE of each estimator's expected change in duration for a one-unit change in a covariate. In both cases smaller values indicate less error, and thus better performance. We run these simulations in R for 1,000 iterations each with sample sizes of 50, 200, 500, and 1,000. Additionally, we vary the amount of right censoring in the data: 5%, 10%, and 20% of the observations at each sample size.

Within each simulation iteration, we obtain N expected durations, but only one marginal change in duration. Each iteration allows us to calculate an RMSE from the N expected durations, but we can only calculate an RMSE for the marginal changes once all of the simulation iterations are complete. Therefore we describe the distribution of 1,000 RMSE statistics for the expected durations, but we can only report one RMSE for marginal changes.

[17] We use the `survival` package in R for model estimation here and in our replication analyses (Therneau 2013).

In Table 1 we present results for the three parametric models and as ratios over the RMSE for COX ED.[18] Ratios that are greater than 1 favor COX ED because the RMSE for COX ED is less than the RMSE for the competitor. The first two columns of results summarize the ratios of the expected duration RMSEs: the average ratio and the proportion greater than 1. The third column of results gives the ratios of the marginal effect RMSEs. All DGPs summarized in Table 1 include 10% right censoring. See the appendix for results with 5% and 20% right censoring.

[Insert Table 1 here]

For each competing model, for both evaluative metrics, and at all four sample sizes, the RMSE is lower for COX ED. The proportion of iterations for which COX ED outperforms the competitor in predicting duration increases with the sample size. This result stems from the fact that as the sample size increases, the empirical shape of the baseline hazard more closely approximates its asymptotic shape implied by the randomly generated baseline hazard. As a result, the parametric models' error increases because the baseline hazard is not drawn from their assumed distributions, but COX ED's performance is not affected because the Cox model makes no such assumption. Likewise, COX ED recovers the true marginal effect with less error than do the parametric models. Furthermore, this improvement increases as the sample size increases to the point that the parametric model's marginal effect error ranges from about 1.70 times to nearly twice as large as the COX ED error with 1,000 observations. Overall, COX ED decisively outperforms the parametric models in accurately generating durations, and in computing the estimated effect of a change to a covariate. Furthermore, the additional results in the appendix shows that this finding holds at smaller and larger proportions of right censoring in the data.

As is mentioned above, we also ran a series of simulations comparing COX ED to the exponential, Weibull, and log-normal survival models after simulating the baseline hazard from those parametric models' assumed distributions. These simulations are described in the appendix. As expected, when the baseline hazard comes from a known distribution (an unlikely situation in ap-

[18] We consider that ratios of RMSE statistics to be more meaningful than the individual RMSEs since the absolute magnitude of RMSE can be influenced by the simulation conditions.

plied research), the corresponding parametric survival model's performance relative to COX ED improves. However, in several instances COX ED still performs nearly equal or better than the parametric models by our RMSE criteria.

7 Applying COX ED to Political Science

Having shown evidence that COX ED performs better than other options for generating expected durations, we next turn to its utility for applied researchers. We re-analyze four published papers that employ the Cox model to assess the extent to which COX ED can help political scientists better understand the substantive implications of their results. These papers span three subfields—comparative politics (Martin and Vanberg 2003), American politics (Box-Steffensmeier 1996; Binder and Maltzman 2002), and international relations (Mattes and Savun 2010). We replicated the Cox model in each article, then employed our Cox ED procedure to assess the substantive effects of key independent variables. Our goal with these replications is not to critique the authors' modeling choices, but rather to demonstrate how COX ED can help applied researchers present Cox model results with more meaningful QI. In doing so, we uncover novel substantive insights in each example. We present the replications of Martin and Vanberg (2003) and Binder and Maltzman (2002) here. To conserve space, we present the Box-Steffensmeier (1996) and Mattes and Savun (2010) replications in the appendix.

7.1 Coalition Bargaining and Government Formation

Martin and Vanberg (2003) examine the determinants of negotiation time among political parties forming a coalition government. In particular, they are interested in the effects of ideological distance between the parties in the coalition as well as the size of the coalition. They use data on government formation in 10 European countries from 1950–1990 to test two hypotheses: that negotiations conclude more quickly (1) when bargaining parties are ideologically close and (2) when there are fewer parties engaged in bargaining (see Martin and Vanberg 2003, 325–326).

The dependent variable in Martin and Vanberg's (2003) analysis is the number of days between the beginning and end of the bargaining period. Martin and Vanberg model this variable as a

function of the *Range of Government*, which is a measure of the ideological distance between the extreme members of the coalition, the *Number of Government Parties* (and its interaction with the natural log of time), and several other variables. Their hypotheses predict negative coefficients on the variables of interest, indicating that increases in the ideological distance between the parties and in the number of parties correspond with a decrease in the risk of government formation, or a longer negotiation time.

The authors demonstrate support for their hypotheses by computing changes in the hazard rate based on changes to these independent variables. Regarding the estimated effect of *Range of Government*, they state the following (331): "an increase in the ideological range of the government from zero (the case of a single-party government) to 1.24 (the average range for coalition governments in our sample) decreases the odds of government formation on any given day in the bargaining process by approximately 23 percent." On the second hypothesis, they find that "for all governments that formed after two weeks of bargaining, negotiations leading to three-party coalitions were on average over 50 percent less likely to end on any particular day than negotiations leading to two-party coalitions" (Martin and Vanberg 2003, 331). Overall, they conclude that both variables are important determinants of the time it takes governments to form.

Martin and Vanberg's (2003) discussion of the substantive effects of their key variables meets the discipline's current standards. However, it also highlights our critiques of relying on the hazard rate. For instance, it is difficult to assess what the estimated effects of *Range of Government* and *Number of Government Parties* mean in substantive terms. How much longer will negotiations take for a typical coalition government than for a single-party government? How long does each additional party delay the process? Our COX ED method is able to answer these kinds of questions.

After replicating the model, we utilized the COX ED method to generate expected durations and confidence intervals for different independent variable profiles. Recall from above that these expected durations are generated from a GAM fit of the observed durations on the expected ranks for each observation produced by the Cox model (Figure 1 displays the GAM fit for this example). The authors expect that as the parties in the coalition become ideologically farther apart and/or

more parties join the coalition, the risk of government formation decreases. Put differently, this means that as *Range of Government* and/or *Number of Government Parties* increases, so too should the expected number of bargaining days. Figure 3 graphs these relationships.

[Insert Figure 3 here]

Panel (a) of Figure 3 graphs the expected number of bargaining days for a comparison that Martin and Vanberg (2003) consider in their analysis: a single-party government (*Range of Government* = 0) versus the average value for coalition governments (*Range of Government* = 1.24). Recall that they report that this change results in an expected 23 percent decrease in the hazard rate. Using COX ED and averaging over the other variables in the model, we estimate about 20 days until government formation for a single party government compared to 22 days for the typical coalition government. This difference is of about 2 days is relatively small and not statistically significant.

Figure 3, panel (b) shows the effect of *Number of Government Parties* with time set to 38 days (its 75th percentile). There we graph the expected increase in bargaining days as a function of three different changes to the number of parties. We estimate that a change from 2 to 3 parties corresponds to an increase of 9 days, moving from 2 to 4 parties lengthens bargaining by 19 days, and moving along the full observed range from 1 to 6 parties makes bargaining longer by almost 1.5 months (43 days). All of these differences are statistically significant at the 0.05 level. We can also examine the impact of time on the effect of coalition size. Panel (c) presents the expected number of bargaining days for 1–6 parties with time set to 1, 19, and 61 days (the 10th percentile, median, and 90th percentile, respectively). The graph shows clear evidence of duration dependence. Adding more parties exerts a substantively small and statistically nonsignificant effect on bargaining time early on, but becomes strongly positive and statistically significant later in the process.

In sum, we find that the effect of *Range of Government* is relatively small compared to the effect of *Number of Government Parties*, at least when a sufficient amount of time has elapsed. The estimated difference of two days due to a change in *Range of Government* is not statistically significant, and even the smallest change in *Number of Government Parties* in panel (b) produces an

estimated effect that is over four times as large. This example illustrates the utility of the COX ED method in assessing Cox model results. While Martin and Vanberg's (2003) analysis of changes in the hazard rate does show that the effect of *Number of Government Parties* is larger than that of *Range of Government*, our analysis adds much more detail about the substantive magnitude of this difference. Moreover, by deriving results in terms of bargaining days, we frame the results in ways that are intuitive and easily understood by a wide audience of readers.

7.2 The Confirmation of Federal Judges

Binder and Maltzman (2002) examine the determinants of the time to confirmation by the Senate for all of the 413 nominees to U.S. Circuit Courts of Appeal between 1947 and 1998. They focus on two types of variables that shape the political context for confirming judges: ideological incentives and institutional opportunities. For example, they expect that the confirmation process takes longer when the opposing party and the president are ideologically far apart and when the nominee would become a critical vote on the court for which he or she is nominated. Institutions also influence the confirmation process. Senators from nominees' home states, for instance, are given the opportunity to object to nominees through the "blue slip" procedure and if the opposing party to the president controls the Senate the majority leader can cause delay in confirmation (see Binder and Maltzman 2002, 191–192).

The dependent variable in this analysis is the number of days a nomination was pending on the Senate floor. Binder and Maltzmann model this variable as a function of several covariates capturing their theoretical framework. In particular, they measure ideological distance between the president and the opposing party in the Senate with DW-NOMINATE scores (*President-Opposing Party Distance*). They also include indicator variables for *Divided Government*, a *Critical Nomination*, an *Ideologically-Distant Home-State Senator*, and a control for whether it is a *Presidential Election Year*. They expect these variables to produce negative coefficients, indicating a decrease in the risk of confirmation, or longer time to confirmation.

The authors demonstrate support for their hypotheses by computing percentage changes in the hazard rate based on changes to these independent variables. Table 2—which reproduces their

Table 3 (196)—reports these estimates. Most notably, all of them are negative (as expected) and statistically significant ($p < 0.05$). From this, Binder and Maltzmann conclude that both ideological and institutional factors influence the duration of the confirmation process.

[Insert Table 2 here]

As in our last example, it is difficult to determine which (if any) of the effects reported in Table 2 are substantively meaningful. For instance, we know that the drop in the hazard rate associated with an *Ideologically-Distant Home-State Senator* is roughly twice the magnitude of *Divided Government* and *Critical Nomination*. But without a meaningful scale, it is difficult to put these factors in context. How do these variables actually influence the amount of time until confirmation? After replicating the model, we utilized COX ED to generate the change in expected durations for the same independent variable profiles shown in Table 2 (see the appendix for the GAM fit). Figure 4 graphs the results.

[Insert Figure 4 here]

Figure 4 shows that for the change in each variable given in Table 2, the expected number of days to confirmation increases. This result is consistent with the decreases in the hazard rate that Binder and Maltzmann report. For example, averaging over the other variables in the model, moving from unified to divided government produces an expected increase in confirmation time of 41 days. Compare that to whether the home-state senator is ideologically close to or distant from the president (during divided government): an increase of about 102 days, averaging over the other variables.

The ability to generate expected durations also allows us to go one step further in assessing the substantive magnitude of these effects. One means of determining whether a confirmation delay is large or small is to consider how many cases a nominee would miss. According to the Federal government's caseload statistics, the median number of filings in the Courts of Appeals in 2000 (the year closest to the authors' data) was 4,069, which amounts to an average of about 11

cases per day.[19] During unified government, a nominee's expected time to confirmation is 46 days according to the COX ED method. This corresponds to absence from 506 cases between the time of nomination and the confirmation date (46 days $\times$ 11 cases/day). When government is divided, that number jumps to 957 (expected duration of 87 days $\times$ 11 cases/day). On average, divided government leads a typical nominee to miss an additional 451 case filings ($957 - 506$) because of delay in the Senate. Now consider the change from an ideologically close to an ideologically distant home-state senator. In the latter scenario the typical nominee misses 1,122 more cases than in the former ($[188 \times 11] - [86 \times 11]$). Clearly the institution of the blue slip gives individual senators substantial control over the confirmation process. Most importantly, this sort of analysis is more substantively meaningful than examining changes in the hazard rate.

8 Conclusions

The Cox model is, for good reason, a popular choice among researchers in political science as well as several other disciplines. The ability to estimate a survival model while leaving the baseline hazard function unspecified makes the Cox model a major contribution to applied statistics. However, this flexibility limits the QI that analysts can compute from their results. Specifically, the typical means of interpreting model results involves multiplicative changes in the hazard rate of event occurrence. This approach may lead researchers to conflate risk and probability, produces substantively vague estimates of covariate effects, and is challenging to effectively communicate to non-academic consumers of research.

As a solution, we present COX ED, a method for computing expected durations from Cox model estimates. The method operates by fitting a GAM of the observed durations to the expected ranks of the observations from the Cox model. This GAM fit is then used to generate expected durations for the observed data or new observations with substantively interesting covariate profiles (which can include time-varying covariates). Importantly, our replications of published studies that employ the Cox model show that these expected durations are useful for substantive discussions of model results. Expected durations are easy to interpret, closely reflect the substantive goal of

[19] See http://www.uscourts.gov/Statistics/FederalJudicialCaseloadStatistics.aspx.

survival analysis (understanding the determinants of the duration of some phenomenon), and can be easily understood by academics, students, journalists, and public officials.

Practically speaking, COX ED is straightforward to implement in the R statistical environment. Our accompanying R package, `coxed`, contains functions that allow researchers to use the method even with minimal knowledge of R syntax. Additionally, the functions are flexible; users can make several changes to many of the features of the method that we describe above. Importantly, the output from the functions provides point estimates, standard errors, and confidence intervals, so researchers can report their results with appropriate measures of uncertainty. In some cases, this uncertainty may be large because the method involves two estimation routines (Cox model and GAM). However, in practice this would only produce conservative, Type II errors rather than lead an analyst to mistakenly find a statistically significant effect. Thus, while it may effectively reduce statistical power in some cases, the considerable benefit to substantive interpretation from COX ED is worthwhile.

Of course, researchers have always had the ability to generate expected durations from a parametric duration model; we do not claim to have developed the quantity for the first time here. However, the parametric models for which expected durations are available force researchers to make an assumption about the baseline hazard function. This assumption may not be correct and is never truly testable. So it is no surprise that the Cox model is a well-used tool in applied work. The main drawback to its popularity is that the substantive clarity of interpretation of results lags behind that of other common statistical models, such as linear regression or logistic regression. COX ED solves this problem. The method provides the benefit of the intuitive QI available in parametric models while retaining the desirable estimation properties of the Cox model.

Appendix

A A Brief Summary of Survival Models

Survival models are designed to provide an explanation for why a particular observation survives for a particular duration of time. A duration, denoted t_i for observation i, can be a patient's lifespan after a diagnosis, the time needed for a negotiation to result in an agreement, or the amount of time that passes before a catastrophic event like government failure or war, among many other examples. While both parametric survival models and the Cox model have similar purposes, they also exhibit key differences. We briefly review these models in a technical discussion below. See Box-Steffensmeier and Jones (2004)—which we rely on extensively for this review—for more details.

A.1 Parametric Survival Models

Survival models improve upon ordinary least squares (OLS) for duration data by allowing for skewed distributions of the durations and by explicitly accounting for the fact that some observations are right censored, meaning that their durations end some time after data collection ends. The likelihood function used by all parametric survival models takes the form

$$L(\theta|\mathbf{t},\mathbf{X}) = \prod_{i=1}^{N} f_i(t)^{\delta_i} \, S_i(t)^{1-\delta_i}, \tag{1}$$

where i indexes observations, θ represents the parameters to be estimated, $\mathbf{t}$ represents the observed durations with t_i referring to the duration of observation i, $\mathbf{X}$ represents the matrix of covariates, N is the sample size, and δ_i is an indicator for the right censored observations. $f_i(t)$ is the PDF of failure times t and

$$S_i(t) = 1 - \int_0^t f_i(t) \, dt \tag{2}$$

is the survivor function, which represents the probability that observation i survives until time t or later.

An important concept in survival modeling is the hazard function, or hazard rate, defined as the ratio of the failure PDF to the survivor function,

$$h_i(t) = \frac{f_i(t)}{S_i(t)}. \tag{3}$$

The hazard rate represents the relative risk of failure at time t_i conditional on survival until time t_i (Box-Steffensmeier and Jones 2004, 15). Results from survival models are often expressed in terms of the hazard ratio, the ratio of two (actual or hypothetical) observations' hazard rates. The failure and survivor functions are different for each observation. These idiosyncratic functions share the same baseline failure PDF, $f_0(t)$, and the variation across cases is induced by the data.

If a parametric survival model can be reparameterized as

$$\log(t_i) = \alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \sigma\varepsilon_i, \tag{4}$$

then the model has an accelerated failure time interpretation in which it is possible to derive the expected duration and marginal change in duration with respect to a covariate. The exponential, Weibull, and log-normal models can all be interpreted in this way. Specifically, in equation 4, $\sigma = 0$ for the exponential model, ε_i is distributed by the type-1 extreme-value distribution for the Weibull model and by the standard normal distribution for the log-normal model (see Box-Steffensmeier and Jones 2004, 23–32).

Since parametric survival models provide an analytic function for the hazard rate, the hazard is assumed to follow one of the paths allowed by this functional form. In many cases this assumption is too restrictive. For instance, the exponential model assumes that the hazard is constant, the Weibull model assumes that the hazard rate is monotonically increasing or decreasing over time, and the log-normal model assumes that the hazard rate is either monotonic or unimodal. Many researchers do not wish to assume that the hazard rate follows exactly one of these forms. As a result, the Cox model has gained wide use in the social sciences relative to the parametric survival models.

A.2 The Cox Proportional Hazards Model

The parameters of the Cox model are estimated by maximizing a partial likelihood function. Cox (1975) shows that the this estimator converges, in the sample size, to the maximum likelihood estimator. Carrying notation from above forward, the partial likelihood is defined as follows

$$\prod_{i=1}^{N} \left[\frac{\exp(\beta' X_i^{(t_i)})}{\sum_{\forall j \in R_{t_i}} \exp(\beta' X_j^{(t_i)})} \right]^{\delta_i} \tag{5}$$

where β is a vector of regression coefficients to be estimated and δ is an indicator for censored observations.

The key feature of the estimator is the term in the denominator: $R(t_i)$. This refers to observation i's "risk set." An observation's risk set is comprised of the observations that experience the event at the same time or after observation i. Thus, the partial likelihood estimator is the product of the conditional probability of failure at a certain time, given all the observations that have not yet failed at that time. This allows the method to estimate parameters while relying only on the ranks of the durations—not the actual durations—and thus avoid making an assumption about the baseline hazard.[20]

However, that advantage comes with some costs. If the baseline hazard assumption is correct, a parametric model will be more efficient than the Cox model because the former uses more information from the data (Box-Steffensmeier and Jones 2004). Moreover, coefficient estimates can only be interpreted with respect to the hazard rate; positive coefficients indicate the hazard is rising while negative estimates signify a decrease. For example, exponentiating a coefficient estimate yields the average multiplicative change in the hazard ratio for a one-unit increase in the independent variable. Similarly, Box-Steffensmeier and Jones (2004, 60) recommend the following formula for computing the percentage change in the hazard rate between two values (X_1 and X_2) of

[20]This also means the partial likelihood estimator is quite sensitive to model specification issues such as omitted variables and measurement error. However, analysts need not abandon the Cox model due to these problems because they can be resolved with a robust estimator of the partial likelihood (see Desmarais and Harden 2012).

an independent variable

$$\%\Delta h(t) = \left[\frac{\exp(\beta[X_i = X_1]) - \exp(\beta[X_i = X_2])}{\exp(\beta[X_i = X_2])}\right] \times 100. \tag{6}$$

As we state above, these hazard rate-based computations are mathematically sound; researchers who employ them are not doing so in error. However, because the Cox model does not use the actual durations, estimates of expected duration are not readily available. In the main text we contend that the interpretation and communication of substantive results from the Cox model can be improved by adapting the Cox model to estimate expected durations and marginal changes in these durations with respect to a covariate. We show that COX ED fulfills this objective.

B Journal Article Meta Analysis

Our meta analysis examined use of the Cox proportional hazards model and methods for interpreting results in four political science journals from 1990–2015: *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, and *International Organization*. We first searched for articles using Google Scholar. Then we coded the articles based on (1) the language used to frame hypotheses and (2) the methods used to interpret Cox model results. We describe the details of these procedures below.

Our central objectives were to assess (1) the type of language researchers typically use to frame their hypotheses when employing the Cox model and (2) the methods they typically use to interpret Cox model results. On the first objective we considered two possible framing styles: a risk frame and a duration frame. A risk frame discusses hypotheses with respect to the risk of event occurrence. For example: "as X increases, the risk of event Y occurring also increases." In such a case the researcher is not primarily concerned with duration, but rather focuses on how the covariates make the event more or less likely to occur. In contrast, a duration frame discusses the hypothesis in terms of event time, as in "as X increases, the number of days until event Y occurs decreases." In this case the length of time that an event takes is of central importance.

B.1 Searching for Articles

We searched https://scholar.google.com/ for "cox proportional hazards" OR "cox model" OR "cox regression" in each journal listed above, one journal at a time. We set the date range to 1990–2015. We downloaded all of the articles returned by these searches, checked each one to make sure it included analysis with the Cox model, then saved it for coding in the next step. We included any paper that reported the estimation of a Cox model in the main text. This produced 63 total articles, ranging in publication date from 1996–2015. The articles came from four subfields: international relations (35), comparative politics (13), American politics (11), and methodology (4). Additionally, the articles spanned all four journals: 20 from *American Journal of Political Science*, 20 from *Journal of Politics*, 12 from *American Political Science Review*, and 11 from *International Organization*.

B.2 Coding Hypothesis Text

First, we identified the number of hypotheses in each article and copied the text of those hypotheses. If multiple analyses were presented, we only included hypotheses pertaining to the Cox model(s) reported in the main text. If no hypotheses were presented with the Cox model (i.e., in a descriptive analysis), we used the authors' descriptions of the model specification (i.e., variables used and purpose of the estimation).

Next, we placed all of the hypotheses' text into a single string, omitted common English stop words, then counted word frequencies of the remaining words. This produced a list of 1,373 unique words, from which we identified words as either predominantly part of a risk frame and words predominantly used in a duration frame.[21] Our general rule was to code any word that related to probability, likelihood, or chance in the risk frame category and any word relating to time in the duration frame category. In all, we coded eight unique words as risk frame words and 47 unique words as duration frame words. Table 3 reports these words and their frequencies.

[Insert Table 3 here]

[21]This was a subjective assessment, and we encourage interested readers to obtain the replication materials and assess whether they agree with our decisions.

B.3 Hypothesis Framing Results

The fact that we coded many more words as duration frame words gives some preliminary evidence that authors tend to frame their hypotheses with respect to the time until event occurrence more often than the risk of event occurrence. However, this may be skewed by the possibility that authors simply have more choices when it comes to duration frame words. Looking at word counts of all eight risk frame words and the top eight duration frame words reveals a larger count of risk frame words: 119 instances of risk frame words and 106 duration frame words. Nonetheless, the full count of all the duration frame words we coded is 215—substantially larger than the total count of risk frame words.

We also coded each article as either predominantly using a risk frame, duration frame, or equal use of both frames. We accomplished this in two ways: a count of unique words and a count of total words. First, we counted how many unique words from each frame appeared in the text of the hypotheses. This approach did not give additional weight to the same word appearing more than once. We then coded an article's frame as the type with the most instances of unique words from its list. Second, for each article we counted the total number of instances of words in its hypotheses from each frame (e.g., allowing for repeats of the same word). In both cases if an equal number of risk and duration frame words appeared, we coded the article as equal.

Using the unique word count, we coded 33 articles as using a duration frame, 16 with a risk frame, and 14 with equal use of both frames. With the total word count these numbers are 33, 20, and 10, respectively. Almost half (29) of the articles use words from both frames, 20 contain no risk frame words, and 14 contain no duration frame words. We also conducted these counts after deleting all of the duration words that appear two or fewer times to check whether these results are driven only by the fact that there may be more choices of duration frame words. In that case the numbers are 25, 18, and 20, further indicating that the duration frame is more common.

B.4 Coding the Interpretation Methods

Our second objective was to code which method(s) each article used to interpret the results of the estimated Cox model. This was accomplished by reading the results sections of the articles and identifying each unique method used. We created a total of four categories based on what we found in the text, which we list below. Note that all of the articles discussed the sign and significance of the Cox model coefficient estimates. The categories reflect any interpretation beyond sign and significance. The articles employed an average of 1.3 of these interpretation methods. 45 articles used one method, 17 articles used two methods, and 1 article employed three different methods.

- *Hazard ratios* (24 articles). This category included any article that reported the exponentiation of one or more Cox model coefficients, as well as a discussion about the resulting multiplicative effect of a one-unit change in the covariate of interest.
- *Changes to the hazard rate* (33 articles). This category included any article that reported a marginal change in the hazard rate (usually expressed as a percentage) corresponding to a substantively interesting change in the values of a covariate.
- *Empirical estimates of the hazard and/or survivor functions* (17 articles). This category included any article that graphically displayed an estimate of the baseline hazard from the model and/or computed the survivor function. Typically this was done for different covariate values to show the effect of changes to the covariate.
- *Only sign and significance of the coefficient estimates* (8 articles). This category included any article that did not report any interpretation of the Cox model other than the sign and significance of the relevant coefficient estimates.

The most important finding from this analysis is the fact that all of the articles that go beyond sign and significance in their interpretation of the Cox model focus on the hazard rate, whether through hazard ratios, changes to the hazard rate, or estimation and graphing of the baseline hazard and/or survivor functions. To further emphasize this point, one article in our data did report expected durations after estimating a Cox model, but those estimates came from re-estimating the model using the Weibull parameterization (Senese and Quackenbush 2003, 714).

B.5 Meta Analysis Conclusions

This analysis yields two important insights. First, we find that political scientists employing the Cox model over the last 25 years tend to discuss their theoretical expectations in the language of time until event occurrence. Language related to the risk of event occurrence also appears, but it is less common than duration-based framing. Approximately 70% of the articles in our sample contain more duration words or an equal amount of duration and risk words (compared to only 48% containing more or equal risk words). It is clear that researchers' substantive interests usually center on the duration of some political phenomenon, not just its likelihood of occurring.

This first finding contrasts sharply with the second finding, which is that researchers nearly exclusively rely on interpretation of the hazard rate after estimating the Cox model. We found no instances where researchers generated expected durations from their Cox model estimates. Thus, researchers who employ the Cox model are typically forced to switch the manner in which they discuss their research when moving from hypotheses to results. This provides motivation for our research, which provides a method for generating expected durations from the Cox model. The COX ED approach allows researchers to maintain consistency between the language they use to describe their theoretical framework and the language they use to communicate their empirical findings.

C The Relationship Between Risk and Failure Probability

Here we show proof that a single hazard ratio almost never corresponds to a single probability ratio. Consider an example of a proportional hazards model in which the coefficients are non-zero. Without loss of generality, consider how an observation t_1 in which $X_1 = 1$ and $X_j = 0$ for $j > 1$ compares to a baseline observation t_0 in which all covariates are zero so that the hazard, failure probability density function (PDF), and survivor function for the observations are all equal to the baseline functions. Let β be the coefficient on X_1. The ratio of the hazard functions for each observation is

$$\frac{h_1(t)}{h_0(t)} = \frac{\exp(\beta)h_0(t)}{h_0(t)} = \exp(\beta). \tag{7}$$

Therefore, a one-unit increase in X_1 is associated with a multiplicative increase of $\exp(\beta)$ in hazard. Now consider how the probability of failure between $t = a$ and $t = b$, conditional on $t > a$, compares for each observation. The conditional probability that the baseline observation fails in this interval is given by

$$Pr(a \leq t_0 \leq b | t_0 > a) = \frac{Pr(a \leq t_0 \leq b)}{Pr(t_0 > a)}. \tag{8}$$

The numerator can be calculated from the baseline failure cumulative distribution function (CDF),

$$Pr(a \leq t_0 \leq b) = \int_a^b f_0(t)\, dt = F_0(b) - F_0(a), \tag{9}$$

and the denominator is the baseline survivor function $S_0(t)$ at $t = a$. The entire conditional probability is given by

$$Pr(a \leq t_0 \leq b | t_0 > a) = \frac{F_0(b) - F_0(a)}{S_0(a)}. \tag{10}$$

Likewise, the conditional probability that the non-baseline observation fails in this interval is

$$Pr(a \leq t_1 \leq b | t_1 > a) = \frac{F_1(b) - F_1(a)}{S_1(a)}. \tag{11}$$

We can rewrite the numerator of (11) as[22]

$$\begin{aligned} F_1(b) - F_1(a) &= [1 - S_1(b)] - [1 - S_1(a)] \\ &= S_1(a) - S_1(b) \\ &= S_0(a)^{\exp(\beta)} - S_0(b)^{\exp(\beta)} \\ &= [1 - F_0(a)^{\exp(\beta)}] - [1 - F_0(b)^{\exp(\beta)}] \\ &= F_0(b)^{\exp(\beta)} - F_0(a)^{\exp(\beta)}, \end{aligned} \tag{12}$$

[22] We exponentiate the coefficient β twice because we raise the baseline survivor function to the power of the hazard ratio, which is $\exp[\beta]$ (see Box-Steffensmeier and Jones 2004, 65, eq. 4.15).

and we can rewrite the denominator of (11) as

$$S_1(a) = S_0(a)^{\exp(\beta)}, \tag{13}$$

so that the conditional probability is

$$Pr(a \leq t_1 \leq b | t_1 > a) = \frac{F_0(b)^{\exp(\beta)} - F_0(a)^{\exp(\beta)}}{S_0(a)^{\exp(\beta)}}. \tag{14}$$

Therefore the ratio of the two probabilities is

$$\frac{Pr(a \leq t_1 \leq b | t_1 > a)}{Pr(a \leq t_0 \leq b | t_0 > a)} = \frac{\frac{F_0(b)^{\exp(\beta)} - F_0(a)^{\exp(\beta)}}{S_0(a)^{\exp(\beta)}}}{\frac{F_0(b) - F_0(a)}{S_0(a)}}$$

$$= \frac{F_0(b)^{\exp(\beta)} - F_0(a)^{\exp(\beta)}}{F_0(b) - F_0(a)} \cdot \frac{S_0(a)}{S_0(a)^{\exp(\beta)}}$$

$$= S_0(a)^{1-\exp(\beta)} \cdot \frac{F_0(b)^{\exp(\beta)} - F_0(a)^{\exp(\beta)}}{F_0(b) - F_0(a)}. \tag{15}$$

In order to consider the multiplicative change in the conditional probability of *instantaneous* failure, let $F_0(b) = x$, $F_0(a) = y$, $S_0(a) = 1 - y$, and $\exp(\beta) = \alpha$, and consider the following limit:

$$\lim_{x \to y} (1-y)^{1-\alpha} \frac{x^\alpha - y^\alpha}{x - y}$$

$$= (1-y)^{1-\alpha} \lim_{x \to y} \frac{x^\alpha - y^\alpha}{x - y}.$$

This limit is the definition of the derivative of the function $g(x) = x^\alpha$ evaluated at $x = y$, so the limit is equal to $g'(y) = \alpha y^{\alpha-1}$. Substituting for y and α, the instantaneous ratio of probabilities is equal to

$$\lim_{b \to a} \frac{Pr(a \leq t_1 \leq b | t_1 > a)}{Pr(a \leq t_0 \leq b | t_0 > a)} = S_0(a)^{1-\exp(\beta)} \exp(\beta) F_0(a)^{\exp(\beta)-1}$$

$$= \exp(\beta) \left(\frac{F_0(a)}{S_0(a)} \right)^{\exp(\beta)-1}. \tag{16}$$

The final expression in (16) is only equal to the hazard ratio, $\exp(\beta)$, when $\beta = 0$ or $F_0(a) = S_0(a)$. The first case is not interesting to an applied researcher because it means the covariate that corresponds to β does not belong in the model. The second case refers to the very specific situation in which a is the median duration, so that the baseline probability of failure before the time point in question is exactly 0.5. For any other duration, the hazard ratio in a proportional hazards model cannot be interpreted as the ratio of conditional failure probabilities at a particular point in time.

To see this illustrated, consider the example of an exponential survival model in which the baseline failure PDF has a mean of 10. Also, suppose we derive a hazard ratio of $\exp(\beta) = 1.1$. This hazard ratio signifies a 10% increase in the risk of failure at time t conditional on survival until time t for a one-unit increase in the covariate. However, in order to find the change in *probability* of failure at time t conditional on survival until time t for a one-unit increase in a covariate, a quantity we denote $w(t)$, we substitute the failure CDF, the survivor function, and $\exp(\beta) = 1.1$ into the formula from (16):

$$w(t) = 1.1\left(\frac{1-\exp[-0.1t]}{\exp[-0.1t]}\right)^{0.1}.$$

At the median of this exponential distribution, $t = 10\ln(2)$, this function evaluates to $w(10\ln[2]) = 1.1$. So the hazard ratio is equal to the probability ratio for the median survival time. But for any other survival time these ratios are different. A hazard ratio of 1.1 implies probability ratios at $t = 4$, $t = 8$, and $t = 12$ of $w(4) = 1.02$, $w(8) = 1.12$, and $w(12) = 1.20$, respectively.

D Simulating Baseline Hazard Functions

Many researchers prefer to use the Cox model in order to avoid making an assumption that the baseline hazard follows a particular functional form. In particular, researchers often do not want to assume that the hazard is constant as in the exponential model, monotonic as in the Weibull model, or unimodal as in the log-normal model. Our challenge in this simulation is to generate durations from functions that represent a variety of more realistic hazard functions that may or may not be monotonic or unimodal. Our method involves fitting a cubic spline to randomly selected points according to the following steps. Figure 5 illustrates key components of the process.

[Insert Figure 5 here]

1. We create a time index that counts integers from 1 to 100. This index serves as the x-axis for the randomly generated baseline hazard function.
2. We draw 10 points on this graph, as illustrated in panel (a) of Figure 5. The x-coordinates for two of the points are 1 and 100, and we randomly draw x-coordinates for the other 8 points without replacement. For example, for the illustration in Figure 5, points are chosen to occur at 1, 6, 12, 20, 40, 51, 55, 71, 85, and 100.
3. We randomly draw the y-coordinates for these points from the standard normal distribution.
4. We fit a cubic smoothing spline to the 10 points. Panel (b) of Figure 5 shows an example.
5. Finally, we apply two transformations to this function to produce a valid PDF. First, we pass the y-values to the standard normal PDF and take the densities. This transformation ensures that the function is non-negative. Second, we divide the y-values by their sum to ensure that the function integrates to 1. This final step in the generation of a baseline hazard function is illustrated in panel (c) of Figure 5.

Having generated a baseline hazard function, our next challenge is to generate individual durations from this function in a way that depends on covariates. To that end, we follow the functional form of the Cox model by using the following steps:

1. We generate a cumulative baseline hazard function by taking the cumulative sum of the baseline hazard.
2. We then create a baseline survivor function from the formula

$$H_0(t) = -\log(S_0[t])$$

   by exponentiating the negative cumulative baseline hazard (Box-Steffensmeier and Jones 2004, 14).
3. We randomly generate three covariates (column vectors of length N denoted X_1, X_2, and X_3), three coefficients (scalars denoted β_1, β_2, and β_3), and a constant (a scalar denoted α) from

the standard normal distribution. We then create a linear predictor $X\beta$ by multiplying

$$X\beta = \begin{bmatrix} 1 & X_1 & X_2 & X_3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

4. We use the baseline survivor function and the linear predictor to construct the individual-specific survivor functions (Box-Steffensmeier and Jones 2004, 65):

$$S_i(t) = S_0(t)^{\exp(\mathbf{X}_i\beta)}.$$

   In other words, we take the baseline survivor function to the power of each element of $\exp(X\beta)$. If, for example, the sample size is 50, then $\exp(X\beta)$ has 50 elements and the baseline survivor function is taken to the power of each of these elements to produce 50 individual-specific survivor functions.
5. We subtract each individual-specific survivor function from 1 and we take the first differences to obtain the individual-specific failure PDFs.
6. In order to draw a duration for each observation from each individual-specific failure PDF we multiply each PDF by 1,000 and round every value up. We then expand a list of integers from 1 to 100 by these rounded values. For example, if after multiplying by 1,000 and rounding up the first two values of the PDF become 5 and 20, then the list of integers from 1 to 100 is expanded to produce 5 copies of 1, 20 copies of 2, and so on. Finally, we draw one randomly selected element from this expanded list. The drawn element becomes the duration for the observation.

Finally, we randomly set a specified proportion of the observations (5%, 10%, or 20%) to be right censored. We then create "true" marginal changes in duration by setting the first covariate in $X\beta$ equal to 1 for every observation, then by setting it equal to 0 for every observation. We pass

these two new linear predictors to the mechanisms we use to generate the durations in each DGP. We then measure the median change in duration.

The simulated durations and the generated covariates together form a simulated dataset. We fit a Cox model (and COX ED) to the simulated data along with an exponential, Weibull, and log-normal survival model. For each model, we compute the expected duration of each observation and compare these estimates to the true durations using an RMSE statistic. To evaluate the estimated marginal effects, we use the fitted models to generate expected durations by setting the first covariate to 1 for every observation and again setting that covariate to 0 for every observation. We subtract the values with $X_1 = 0$ from those with $X_1 = 1$ and save the median of the differences. We then compare those medians to the true marginal effects using another RMSE statistic.

E Simulations with 5% and 20% Right Censoring

Table 1 in the main text reports simulation results from the random spline DGP with the proportion of right censoring fixed at 10% of the observations. In Tables 4 and 5 we report results with 5% and 20% right censoring, respectively. As before, we present results for the three parametric models as ratios over the RMSE for COX ED. Ratios that are greater than 1 favor COX ED because the RMSE for COX ED is less than the RMSE for the competitor. The first two columns of results summarize the ratios of the expected duration RMSEs: the average ratio and the proportion greater than 1. The third column of results gives the ratios of the marginal effect RMSEs.

[Insert Table 4 here]

[Insert Table 5 here]

As with the results in the main text, we find that the RMSE is lower for COX ED across the different simulation conditions. The proportion of iterations for which COX ED outperforms the competitor in predicting duration again increases with the sample size. Additionally, COX ED recovers the true marginal effect with less error than do the parametric models. Furthermore, this improvement again increases as the sample size increases *and* increases as the amount of

right censoring increases. Overall, we find that COX ED is superior to the parametric models in accurately generating durations, and in computing the estimated effect of a change to a covariate.

F Simulations from Parametric Hazard Functions

In the simulation described in section 6 and the appendix, we generate simulated durations from baseline hazard functions that do not follow any particular functional parametric form. We argue that these baseline hazard functions are more realistic than common parametric functions. However, these functions may also favor COX ED over the competing survival models because the exponential, Weibull, and log-normal models are misspecified. To compare the estimators under ideal conditions for the parametric models, we simulate the durations from the assumed distribution of each of the parametric models, then compare each model to COX ED.

We conduct three simulations from parametric hazard functions. First, we generate the baseline hazard from an exponential distribution in which the rate parameter is set to $\frac{1}{\exp(X\beta)}$, and we compare the relative performance of the exponential survival model and COX ED. Second, we generate the baseline hazard from a Weibull distribution in which the scale parameter is set to $\exp(X\beta)$ and the shape parameter to 5, and we consider the performance of COX ED relative to the Weibull survival model. Finally, we generate the baseline hazard from the log of the normal distribution with a mean equal to $\exp(X\beta)$ and a standard deviation equal to 1, and we compare COX ED and the log-normal survival model. In all of the parametric simulations we set the proportion of right censoring to 10% of the observations. Table 6 presents the parametric simulation results. It follows the format of Table 1 in section 6 of the main text.

[Insert Table 6 here]

The expected duration RMSEs show that COX ED and the exponential model are roughly similar in performance when the DGP is exponential. The expected durations slightly favor COX ED, while the marginal effect RMSE ratios indicate that the exponential model is somewhat better. The Weibull model results show a stronger pattern. When the true DGP comes from the Weibull distribution, the Weibull model generally outperforms COX ED with respect to recovering expected

durations and marginal changes in duration. The log-normal results are somewhat different. In that case COX ED produces the smaller expected duration RMSEs across the four sample sizes. However, the marginal effect RMSE ratios show that the log-normal model outperforms COX ED at the sample sizes larger than 50.

Overall, these results show that when the baseline hazard comes from a known distribution (an unlikely situation in applied research), the corresponding parametric survival model's performance relative to COX ED improves. However, in several instances COX ED still performs nearly equal or better than the parametric models by our RMSE criteria.

G Box-Steffensmeier (1996) Replication

Box-Steffensmeier (1996) examines whether U.S. House incumbents' ability to raise campaign funds can effectively deter quality challengers from entering the race. The theoretical expectation is that as incumbents raise more money, challengers further delay their decision to run for the incumbent's seat. She employs data on 397 House races in the 1989–1990 election cycle to test this hypothesis.

The dependent variable in this analysis is the number of weeks after January 1, 1989 when a challenger entered the race. Races in which no challenger entered are coded as the number of weeks after January 1 when the state's primary filing deadline occurred, and are treated as censored. The key independent variable is the incumbent's *War Chest*, or the amount of money in millions of dollars that the incumbent has in reserve at a given time. Importantly, this measure updates over the course of five Federal Election Commission (FEC) reporting periods, so it is a time-varying covariate. The theory predicts a negative coefficient on this variable, which would indicate that as the incumbent raises more money, the hazard of challenger entry declines (and the time until entry increases).

The results of the Cox model provide support. The coefficient on *War Chest* is negative and statistically significant. Box-Steffensmeier explains that "each $100,000 in an incumbent's war chest decreases the hazard of a high quality challenger entering by 16%" (365). Thus, the data indicate that building a war chest is an effective way to avoid being challenged in an election.

We employed COX ED to give an interpretation of these results in terms of the number of weeks a challenger's entry is expected to be delayed with a change in the incumbent's fundraising efforts. We contend that this is more meaningful than the hazard-rate QI that Box-Steffensmeier reports. Additionally, this re-analysis provides an example of the use of COX ED with a time-varying covariate.[23] Figure 6 gives the expected duration, in weeks, until challenger entry for two values of *War Chest*: $710,000 (the median) and $850,000 (the 80th percentile). It also reports the difference between those two estimates.

[Insert Figure 6 here]

Figure 6 supports Box-Steffensmeier's (1996) assertion that the size of an incumbent's campaign funds corresponds with a delay in challenger entry. All else equal, an incumbent with $710,000 in reserve expects to face a challenger 51 weeks from January 1 while one with $850,000 in campaign money will not be challenged until 59 weeks. However, it is important to note that there is quite a bit of uncertainty around these estimates and so the difference of 8 weeks is not statistically significant.

The relatively large confidence intervals shown in Figure 6 appear due to the fact that only a small number of challengers appear in the data, and so many observations are right censored. As a result, the GAM is fit with only 40 observations (see Figure 8). This highlights a potential drawback of COX ED. The GAM relies on the non-censored data, and so it will be estimated with more uncertainty if a large proportion of the observations are censored. However, because the method accounts for uncertainty from the Cox model and the GAM, this makes it susceptible to the more conservative Type II errors: failing to find a significant effect when in truth there is one. In this case, while not statistically significant, the substantive magnitude of an 8-week difference is still noteworthy. A delay of two months over the course of a campaign gives an incumbent a considerable amount of time to generate electoral support without competition.

[23] A function called `cox.ed.tvc()` in our `coxed` R package can perform the COX ED procedure using the counting-process data structure that time-varying covariates require.

H Mattes and Savun (2010) Replication

Our final replication study is Mattes and Savun's (2010) analysis of the duration of civil war peace agreements. The central point the authors make is that provisions that require parties to reveal otherwise private military information can greatly increase the endurance of an agreement. Using data covering 51 civil wars from 1945–2005, they quantify the effect of peace agreements with provisions designed to reduce uncertainty between sides on the life of the agreement. These provisions include third-party monitoring, encouraging belligerents to provide troop and weapon information, and third-party verification of information (see Mattes and Savun 2010, 516–517).

The dependent variable is the number of months a peace agreement lasted. Mattes and Savun (2010) model this variable as a function of several covariates: a count of the *Uncertainty-Reducing Provisions* in the peace agreement and control variables. They hypothesize that "[t]he greater the number of uncertainty-reducing provisions in a civil war agreement, the less likely is the recurrence of civil war between domestic belligerents" (517). This hypothesis predicts a negative coefficient on *Uncertainty-Reducing Provisions*, indicating that as the number of provisions increases, the hazard of peace failure declines (longer peace times).

The Cox model results support the authors' hypothesis, producing a negative and statistically significant estimate on *Uncertainty-Reducing Provisions*. Mattes and Savun (2010) report that its effect is "not only statistically significant but also substantively important" (521). An increase from zero provisions to one provision corresponds with a 46% drop in the hazard rate of peace failure and an increase from zero to three provisions decreases the hazard rate by 84%. From this, they conclude that provisions that reveal information about warring parties are a useful policy prescription for the international community.

We use COX ED to better understand the implications of the results for future peace agreements. The authors label a drop of 46% in the hazard rate as "substantively important." This leads to a key question: what percentage drop would be considered *not* substantively important? Would 10% or 20% be too small to indicate that *Uncertainty-Reducing Provisions* exerts a meaningful effect? Assessing the magnitude of effects is always arbitrary to some degree, but this issue is

compounded when the scale of the effect is not meaningful. It is difficult to state whether a drop of 46% really is "large" or "small." Using our COX ED method, we assess the impact of *Uncertainty-Reducing Provisions* on a much more intuitive quantity: the amount of time peace is expected to last.

Figure 7 supports Mattes and Savun's (2010) assertion that *Uncertainty-Reducing Provisions* exerts a substantively important effect on the duration of civil war peace agreements, though there is a great deal of uncertainty in the estimates. Averaging over the rest of the model, an agreement with no provisions is expected to last about 89 months. Including one provision increases that estimate to about 109, or a gain of 20 months. Moving to two and three provisions brings the estimate to 126 and 158 months, respectively. However, while all of these estimates are statistically significantly different from zero, they are not statistically distinguishable from one another. This is not too surprising given the small sample of 51 cases. More importantly, the data suggest that these estimates are substantively meaningful. The expected difference between a case with no provisions and one with three provisions is 69 months, or the equivalent of moving from the 25^{th} percentile of the observed durations to the 55^{th}. Put differently, it represents almost six additional years of peace. Despite the large confidence intervals, these results indicate that provisions that reduce uncertainty play an important role in the life of peace agreements.

[Insert Figure 7 here]

While we reach the same general conclusion as do Mattes and Savun (2010), our analysis using COX ED provides more substantive detail on the effects of *Uncertainty-Reducing Provisions* on civil war peace duration. This is particularly important given that the authors' research carries important policy implications. They state that "[e]ncouraging the adoption of such uncertainty-reducing provisions in civil war settlements may be a useful policy in the international community's effort to establish peace in civil-war-torn societies" (512). We suspect that should political scientists be given the forum to formally make such recommendations, presenting evidence in terms of the expected length of peace agreements rather than relative changes in the hazard rate would be more intuitive to and make a stronger impression on policymakers.

I Replication Model GAM Fits

Figure 8 presents the COX ED GAM fits for the Binder and Maltzman (2002), Box-Steffensmeier (1996), and Mattes and Savun (2010) replication models. In all three graphs the points represent non-censored observations, which are used to fit the GAMs.

[Insert Figure 8 here]

References

Beck, Nathaniel, and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42(2): 596–627.

Bender, Ralf, Thomas Augustin, and Maria Blettner. 2005. "Generating Survival Times to Simulate Cox Proportional Hazards Models." *Statistics in Medicine* 24(11): 1713–1723.

Berliner, Daniel, and Aaron Erlich. 2015. "Competing for Transparency: Political Competition and Institutional Reform in Mexican States." *American Political Science Review* 109(1): 110–128.

Binder, Sarah A., and Forrest Maltzman. 2002. "Senatorial Delay in Confirming Federal Judges, 1947–98." *American Journal of Political Science* 46(1): 190–199.

Box-Steffensmeier, Janet M. 1996. "A Dynamic Analysis of the Role of War Chests in Campaign Strategy." *American Journal of Political Science* 40(2): 352–371.

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

Box-Steffensmeier, Janet M., Laura W. Arnold, and Christopher J. W. Zorn. 1997. "The Strategic Timing of Position Taking in Congress: A Study of the North American Free Trade Agreement." *American Political Science Review* 91(2): 324–338.

Carsey, Thomas M., and Jeffrey J. Harden. 2013. *Monte Carlo Simulation and Resampling Methods for Social Science*. Thousand Oaks, CA: Sage.

Collett, David. 2003. *Modelling Survival Data in Medical Research, Second Edition*. Boca Raton, FL: Chapman & Hall/CRC.

Cox, Christopher, Haitao Chu, Michael F. Schneider, and Alvaro Munoz. 2007. "Parametric Survival Analysis and Taxonomy of Hazard Functions for the Generalized Gamma Distribution." *Statistics in Medicine* 26(23): 4352–4374.

Cox, David R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society. Series B (Methodological)* 34(2): 187–220.

Cox, David R. 1975. "Partial Likelihood." *Biometrika* 62(2): 269–276.

Cox, David R., and David Oakes. 1984. *Analysis of Survival Data*. Monographs on Statistics & Applied Probability Boca Raton, FL: Chapman & Hall/CRC.

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.

Diermeier, Daniel, and Peter van Roozendaal. 1998. "The Duration of Cabinet Formation Processes in Western Multi-Party Democracies." *British Journal of Political Science* 28(4): 609–626.

Fortna, Virginia Page. 2004. "Does Peacekeeping Keep Peace? International Intervention and the Duration of Peace After Civil War." *International Studies Quarterly* 48(2): 269–292.

Gandrud, Christopher. 2015. "simPH: An R Package for Illustrating Estimates for Interactive and Nonlinear Effects from Cox Proportional Hazard Models." Forthcoming, *Journal of Statistical Software*.

Hanmer, Michael J., and Kerem Ozan Kalkan. 2013. "Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models." *American Journal of Political Science* 57(1): 263–277.

Harden, Jeffrey J. 2011. "A Bootstrap Method for Conducting Statistical Inference with Clustered Data." *State Politics & Policy Quarterly* 11(2): 223–246.

Hernan, Miguel A. 2010. "The Hazards of Hazard Ratios." *Epidemiology* 21(1): 13–15.

Imai, Kosuke, Gary King, and Olivia Lau. 2012. *Zelig: Everyone's Statistical Software*. R Package: version 3.5.4.

Kalbfleisch, John D., and Ross L. Prentice. 2002. *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: Wiley-Interscience.

Katz, Jonathan N., and Brian R. Sala. 1996. "Careerism, Committee Assignments, and the Electoral Connection." *American Political Science Review* 90(1): 21–33.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2): 341–355.

Krustev, Valentin L. 2006. "Interdependence and the Duration of Militarized Conflict." *Journal of Peace Research* 43(3): 243–260.

Lo, Nigel, Barry Hashimoto, and Dan Reiter. 2008. "Ensuring Peace: Foreign-Imposed Regime Change and Postwar Peace Duration, 1914–2001." *International Organization* 62(4): 717–736.

Maltzman, Forrest, and Charles R. Shipan. 2008. "Change, Continuity, and the Evolution of Law." *American Journal of Political Science* 52(2): 252–267.

Martin, Lanny W., and Georg Vanberg. 2003. "Wasting Time? The Impact of Ideology and Size on Delay in Coalition Formation." *British Journal of Political Science* 33(2): 323–344.

Mattes, Michaela, and Burcu Savun. 2010. "Information, Agreement Design, and the Durability of Civil War Settlements." *American Journal of Political Science* 54(2): 511–524.

Meernik, James, and Chelsea Brown. 2007. "The Short Path and the Long Road: Explaining the Duration of U.S. Military Operations." *Journal of Peace Research* 44(1): 65–80.

Senese, Paul D., and Stephen L. Quackenbush. 2003. "Sowing the Seeds of Conflict:The Effect of Dispute Settlements on Durations of Peace." *Journal of Politics* 65(3): 696–717.

Shipan, Charles R., and Megan L. Shannon. 2003. "Delaying Justice(s): A Duration Analysis of Supreme Court Confirmation." *American Journal of Political Science* 47(4): 654–668.

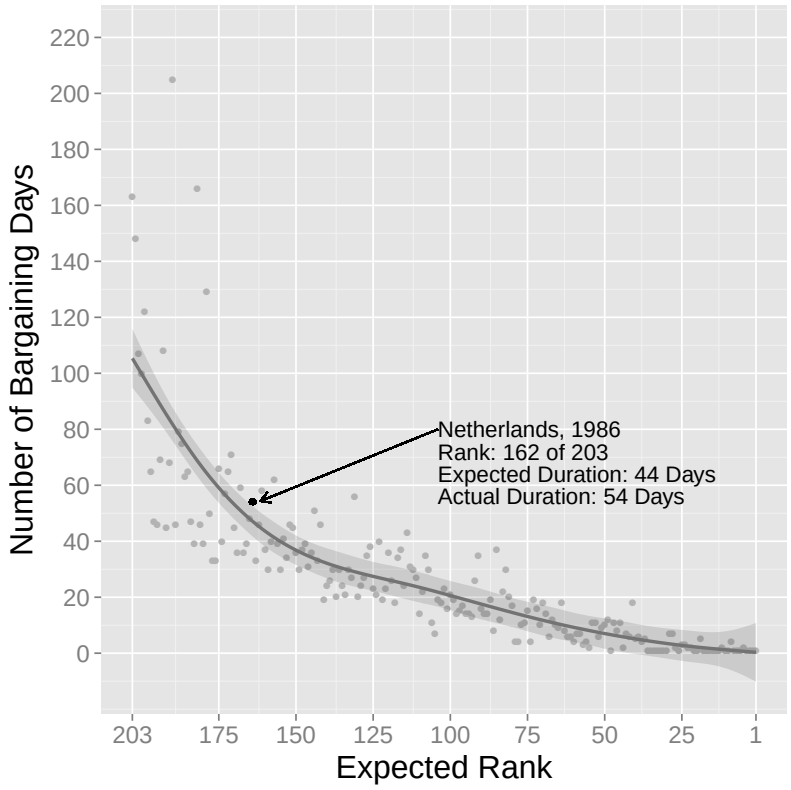
Therneau, Terry. 2013. "`survival`: A Package for Survival Analysis in S." R package version 2.37-4. http://CRAN.R-project.org/package=survival.

Uno, Hajime, et al. 2014. "Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis." *Journal of Clinical Oncology* 32(22): 2380–2385.

Wood, Simon N. 2006. *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

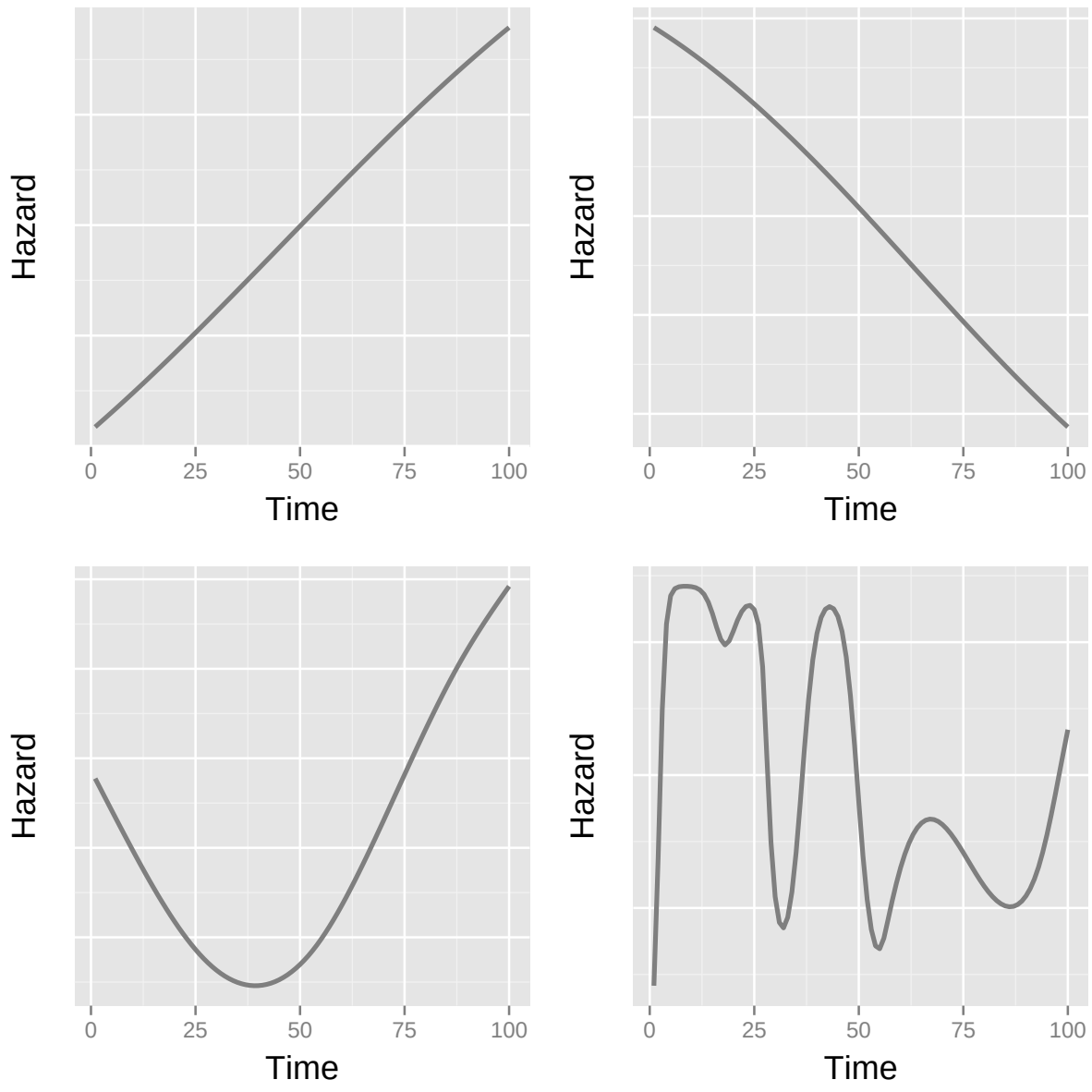
Wood, Simon N. 2011. "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." *Journal of the Royal Statistical Society. Series B (Methodological)* 73(1): 3–36.

Figure 1: GAM Fit of the Observed Durations Against Expected Ranks from the Martin and Vanberg (2003) Cox Model



Note: The graph plots the expected ranks of the observations on the *x*-axis (descending order) against the observed durations on the *y*-axis. The solid line and shading indicate the GAM fit and its 95% confidence interval.

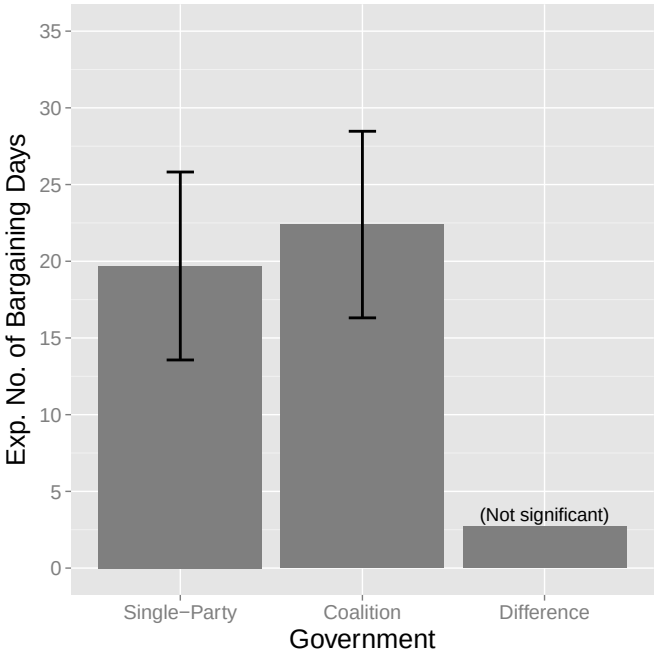
Figure 2: Examples of Baseline Hazard Functions Generated with the Random Spline Method



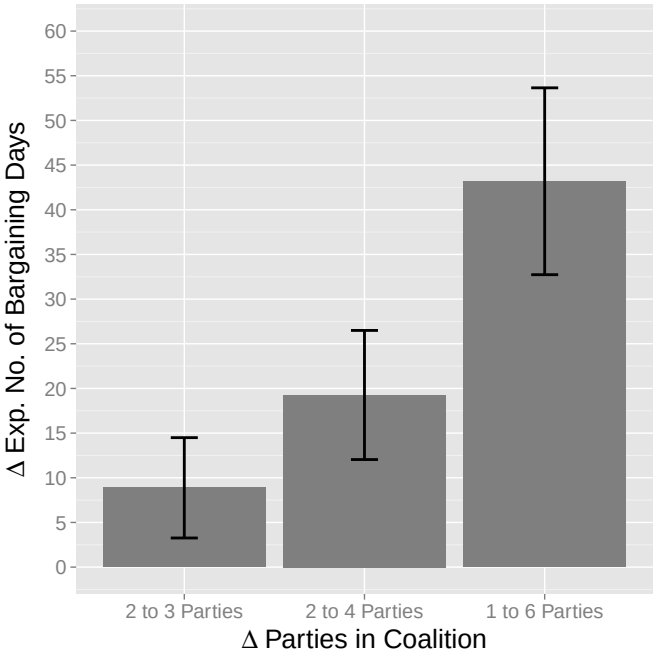
Note: The graphs illustrate four examples of baseline hazard functions generated with the random spline method for use in the simulations.

Figure 3: The Effects of *Range of Government* and *Number of Government Parties* on the Expected Number of Bargaining Days until Coalition Government Formation (Martin and Vanberg 2003)

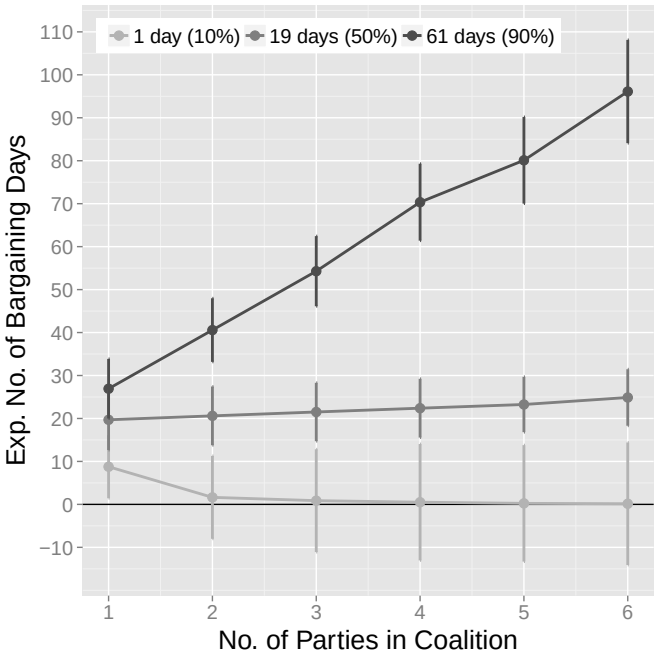
(a) *Range of Government*



(b) Δ *Parties in Coalition (Time = 38 Days)*

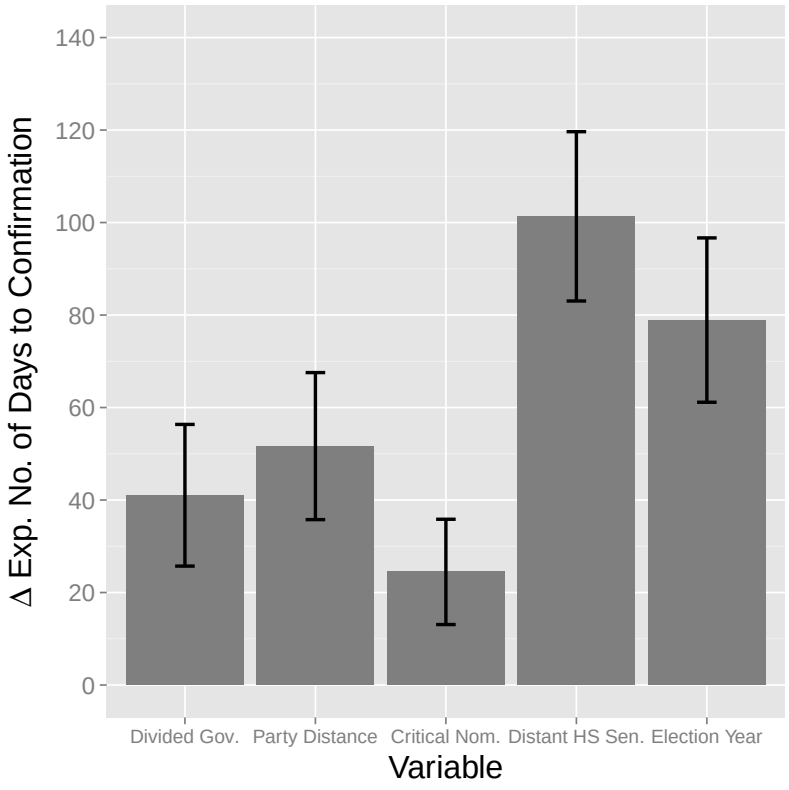


(c) *Parties in Coalition (Time = 1, 19, and 61 Days)*



Note: Panel (a) graphs the expected number of bargaining days for a single-party government (*Range of Government* = 0) versus the average value for coalition governments (*Range of Government* = 1.24) and the difference between the two. Panel (b) graphs the increase in the expected number of bargaining days that corresponds to three different changes in the number of parties in the coalition and time set to 38 days (75 th percentile). Panel (c) graphs the expected number of bargaining days for 1–6 parties with time set to 1, 19, and 61 days (the 10th percentile, median, and 90th percentile, respectively). Brackets and vertical lines indicate 95% confidence intervals.

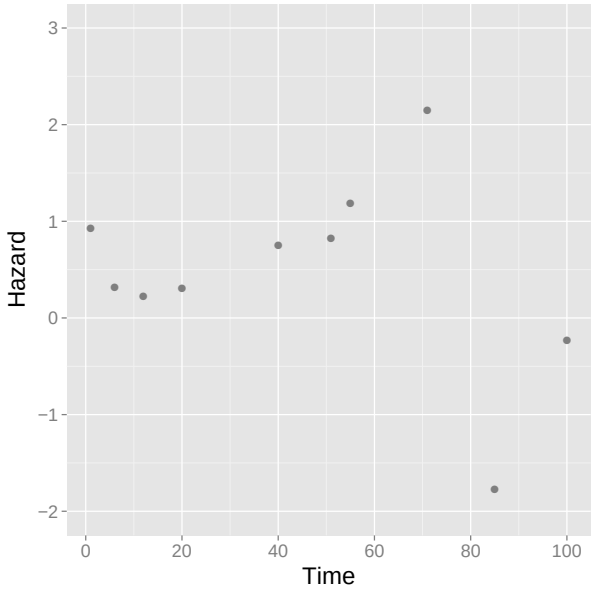
Figure 4: The Determinants of the Time to Confirmation on U.S. Circuit Courts of Appeal (Binder and Maltzman 2002)



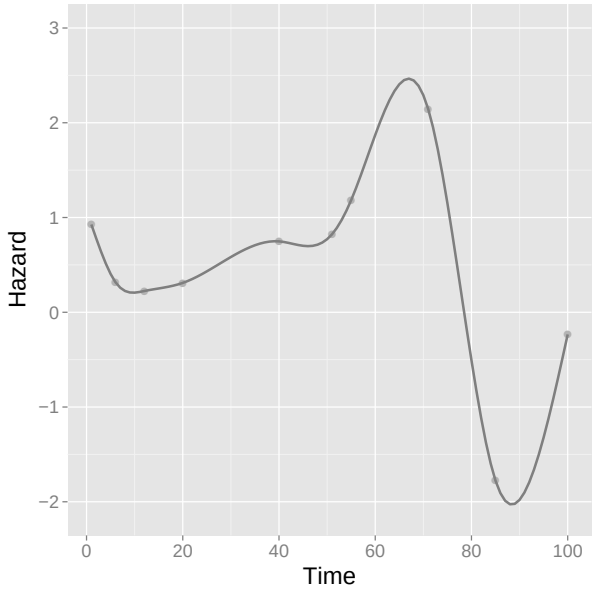
Note: The graph plots the change in the expected number of days to confirmation corresponding to the changes in each variable listed in Table 2. Brackets indicate 95% confidence intervals.

Figure 5: An Example of the Random Spline DGP

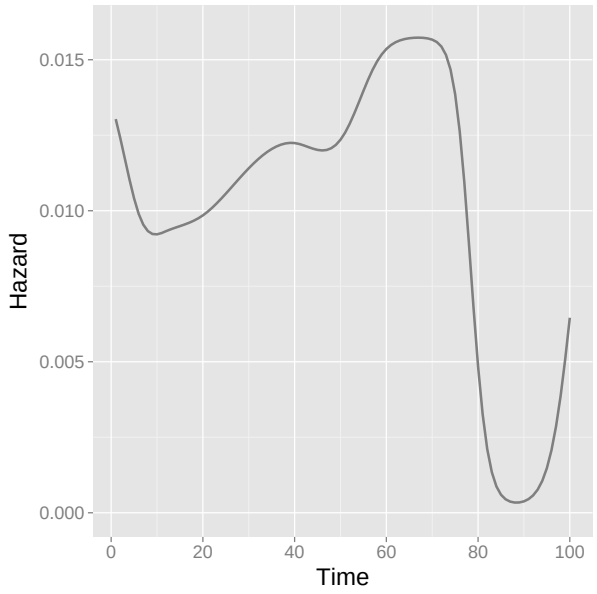
(a) Graph Points



(b) Cubic Spline Fit

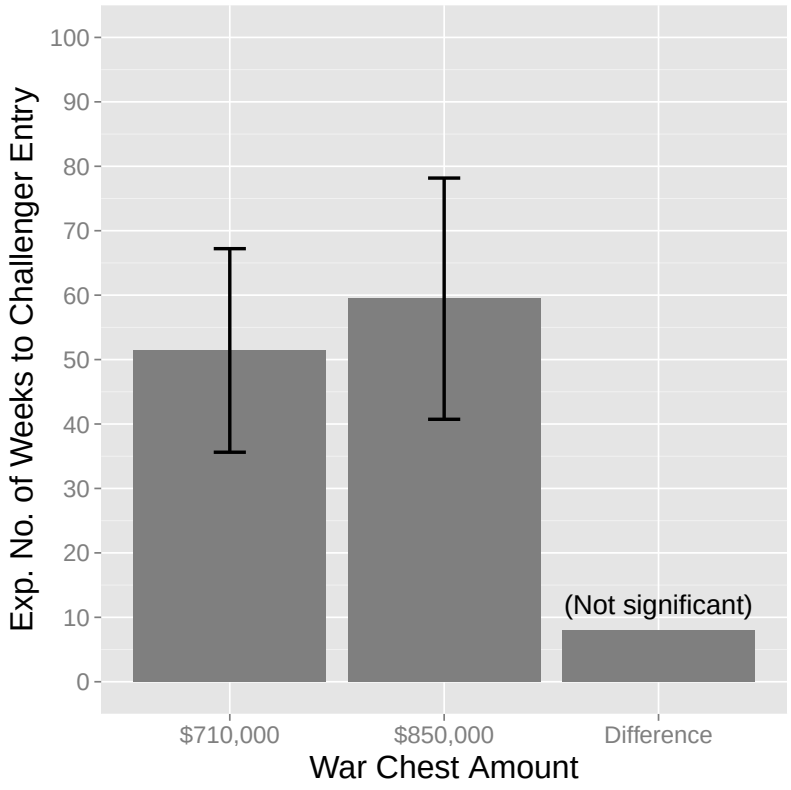


(c) Transform to PDF



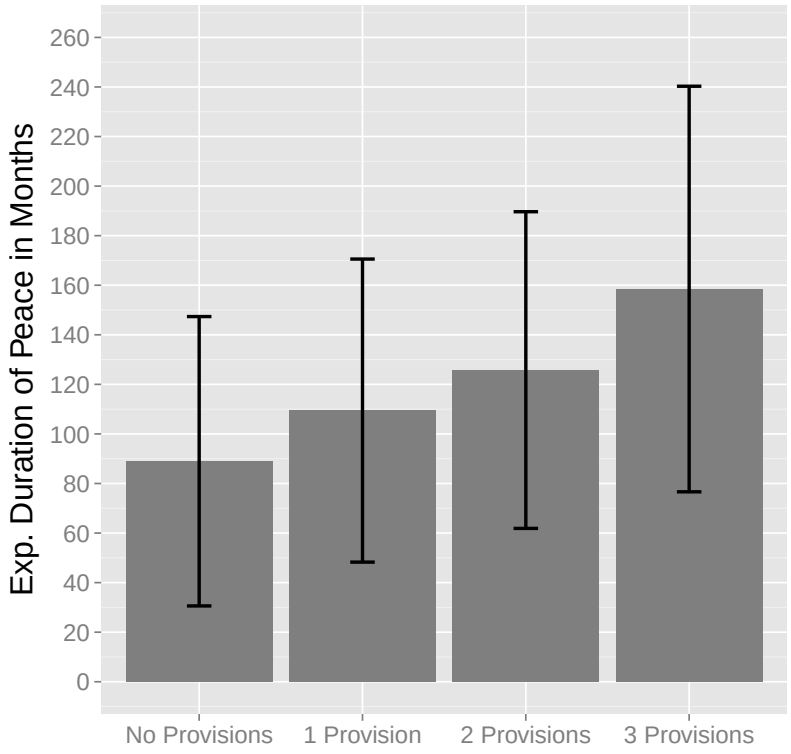
Note: Panel (a) shows an example of 10 randomly-drawn time points (steps #1–3). Panel (b) gives the cubic spline fit to those points (step #4). Panel (c) shows the transformation from the cubic spline to a valid PDF (step #5).

Figure 6: The Effect of Incumbent *War Chest* on the Expected Time Until Quality Challenger Entry into U.S. House Races (Box-Steffensmeier 1996)



Note: The graph plots the expected number of weeks until quality challenger entry for the two values of an incumbent's war chest and the difference between the two estimates. Brackets indicate 95% confidence intervals.

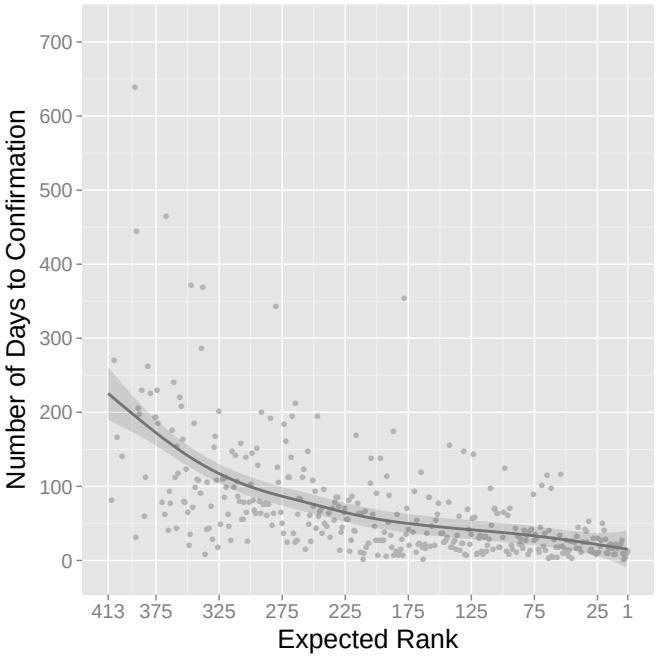
Figure 7: The Effect of *Uncertainty-Reducing Provisions* on the Expected Duration of Civil War Peace Agreements (Mattes and Savun 2010)



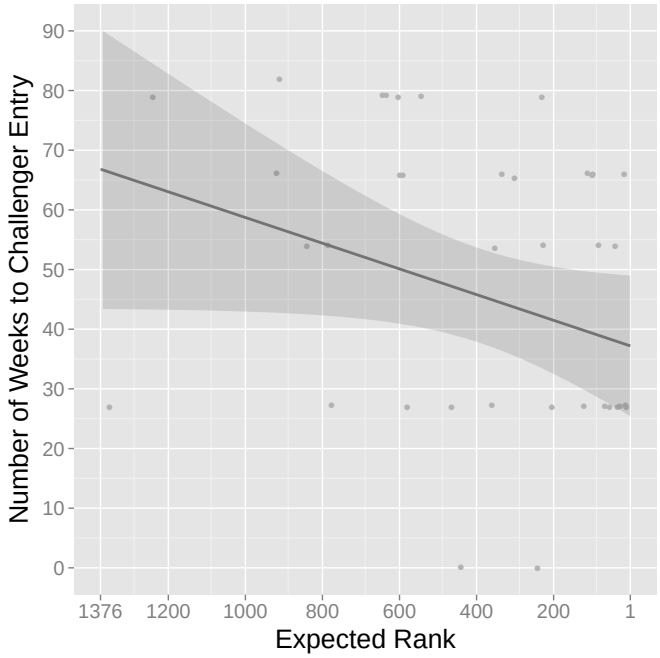
Note: The graph plots the expected peace agreement length, in months, for each observed value of *Uncertainty-Reducing Provisions*. Brackets indicate 95% confidence intervals.

Figure 8: Replication Model GAM Fits

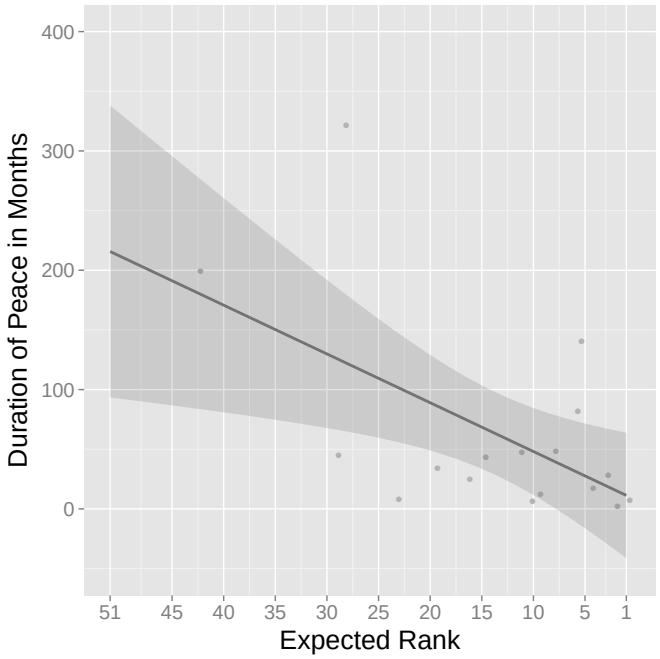
(a) Binder and Maltzman (2002)



(b) Box-Steffensmeier (1996)



(c) Mattes and Savun (2010)



Note: The graphs present the COX ED GAM fits for the Binder and Maltzman (2002), Box-Steffensmeier (1996), and Mattes and Savun (2010) models. Shading indicates 95% confidence intervals.

Table 1: Comparison of Expected Duration RMSE and Marginal Change in Expected Duration RMSE with the Random Spline DGP and 10% Right Censoring

Model	Sample Size	Expected Durations		Marginal Δ
		Average Ratio	% Ratios > 1	Ratio
Exponential	50	1.234	94%	1.178
	200	1.170	98%	1.555
	500	1.156	99%	1.740
	1,000	1.154	99%	1.932
Weibull	50	1.228	97%	1.053
	200	1.175	99%	1.408
	500	1.165	100%	1.525
	1,000	1.160	100%	1.696
Log-normal	50	1.183	96%	1.142
	200	1.145	100%	1.459
	500	1.133	100%	1.712
	1,000	1.136	100%	1.849

Note: Cell entries report the ratio of the parametric models' RMSEs to COX ED's RMSE for each model/sample size combination. Values greater than 1 indicate better performance by COX ED. The first two columns of results summarize the ratios of the expected duration RMSEs: the average ratio and the proportion greater than 1. The third column of results gives the ratios of the marginal effect RMSEs. The proportion of observations that are right censored is fixed at 10%.

Table 2: Estimated Hazard Rate Changes for Key Variables (Binder and Maltzman 2002, Table 3, 196)

Variable	Change in Variable	% Change in Hazard Rate	95% Conf. Interval
Divided Government	No to Yes	−46.51	[−49.29, −46.15]
President-Opposing Party Distance	Low to High	−60	[−65, −54.73]
Critical Nomination during *Divided Government*	No to Yes	−42.22	[−54.44, −27.79]
Ideologically-Distant Home-State Senator during *Divided Government*	No to Yes	−92.44	[−95.67, −87.28]
Presidential Election Year	No to Yes	−73.96	[−75.13, −72.76]

Table 3: Frequency and Frame of Hypothesis Framing Words

Word	Frequency	Coded Frame
likely	73	Risk
risk	16	Risk
likelihood	8	Risk
probability	8	Risk
hazard	7	Risk
hazards	4	Risk
propensity	2	Risk
odds	1	Risk
time	26	Duration
timing	14	Duration
duration	13	Duration
longer	12	Duration
end	11	Duration
early	10	Duration
following	10	Duration
stable	10	Duration
delay	9	Duration
tenure	9	Duration
delays	8	Duration
earlier	8	Duration
termination	8	Duration
future	8	Duration
deadline	7	Duration
finite	5	Duration
quickly	4	Duration
survival	4	Duration
durable	3	Duration
short	3	Duration
date	2	Duration
deadlines	2	Duration
live	2	Duration
long	2	Duration
term	2	Duration
terminate	2	Duration
conclude	1	Duration
concludes	1	Duration
consistently	1	Duration
delayed	1	Duration
delaying	1	Duration
durability	1	Duration
immediately	1	Duration
indefinitely	1	Duration
last	1	Duration
lasts	1	Duration
longer	1	Duration
longer lasting	1	Duration
resilient	1	Duration
retards	1	Duration
shorter	1	Duration
shortrun	1	Duration
slow	1	Duration
slowly	1	Duration
onset	1	Duration
survive	1	Duration
temporal	1	Duration

Table 4: Comparison of Expected Duration RMSE and Marginal Change in Expected Duration RMSE with the Random Spline DGP and 5% Right Censoring

Model	Sample Size	Expected Durations		Marginal Δ
		Average Ratio	% Ratios > 1	Ratio
Exponential	50	1.179	87%	1.125
	200	1.129	95%	1.515
	500	1.119	96%	1.602
	1,000	1.116	98%	1.807
Weibull	50	1.197	96%	1.041
	200	1.145	99%	1.391
	500	1.136	99%	1.527
	1,000	1.132	99%	1.617
Log-normal	50	1.191	99%	1.118
	200	1.152	100%	1.403
	500	1.141	100%	1.685
	1,000	1.143	100%	1.768

Note: Cell entries report the ratio of the parametric models' RMSEs to COX ED's RMSE for each model/sample size combination. Values greater than 1 indicate better performance by COX ED. The first two columns of results summarize the ratios of the expected duration RMSEs: the average ratio and the proportion greater than 1. The third column of results gives the ratios of the marginal effect RMSEs. The proportion of observations that are right censored is fixed at 5%.

Table 5: Comparison of Expected Duration RMSE and Marginal Change in Expected Duration RMSE with the Random Spline DGP and 20% Right Censoring

Model	Sample Size	Expected Durations		Marginal Δ
		Average Ratio	% Ratios > 1	Ratio
Exponential	50	1.441	99%	1.332
	200	1.316	99%	1.796
	500	1.290	100%	1.947
	1,000	1.290	100%	2.266
Weibull	50	1.302	99%	1.170
	200	1.261	99%	1.538
	500	1.240	100%	1.766
	1,000	1.234	100%	1.934
Log-normal	50	1.202	91%	1.208
	200	1.152	97%	1.563
	500	1.134	99%	1.733
	1,000	1.136	99%	2.027

Note: Cell entries report the ratio of the parametric models' RMSEs to COX ED's RMSE for each model/sample size combination. Values greater than 1 indicate better performance by COX ED. The first two columns of results summarize the ratios of the expected duration RMSEs: the average ratio and the proportion greater than 1. The third column of results gives the ratios of the marginal effect RMSEs. The proportion of observations that are right censored is fixed at 20%.

Table 6: Comparison of Expected Duration RMSE and Marginal Change in Expected Duration RMSE with the Parametric DGPs

Model	Sample Size	Expected Durations		Marginal Δ
		Average Ratio	% Ratios > 1	Ratio
Exponential	50	1.167	77%	0.890
	200	1.032	67%	0.830
	500	1.023	65%	0.915
	1,000	0.997	58%	0.912
Weibull	50	1.093	67%	0.893
	200	0.945	53%	0.443
	500	0.896	44%	0.537
	1,000	0.879	42%	0.405
Log-normal	50	1.145	96%	1.058
	200	1.086	100%	0.949
	500	1.066	100%	0.243
	1,000	1.057	100%	0.208

Note: Cell entries report the ratio of the parametric models' RMSEs to COX ED's RMSE for each model/sample size combination with the parametric DGPs. Values less than 1 indicate better performance by the parametric models. Values greater than 1 indicate better performance by COX ED. The first two columns of results summarize the ratios of the expected duration RMSEs: the average ratio and the proportion greater than 1. The third column of results gives the ratios of the marginal effect RMSEs. The proportion of observations that are right censored is fixed at 10%.