Asking About 'Which': Improving Substantive Interpretations of Duration Models[*]

Shawna K. Metzger
National University of Singapore
smetzger@nus.edu.sg

Benjamin T. Jones
University of Mississippi
btjones1@olemiss.edu

March 13, 2017

Abstract:
Despite duration models' use in political science, no best practice exists for substantively interpreting their results. Many practitioners simply stop after interpreting the model's estimated coefficients. When substantive interpretation does happen, practitioners use a variety of strategies. Interpreting the hazard rate is among the more prevalent strategies, particularly for Cox models. Yet, hazards are not intuitive quantities with an easily interpretable meaning for many. We offer a new interpretation technique that removes the need to directly interpret hazards: transition probabilities, which are a more intuitive quantity for readers to engage with, in general. We show how to compute transition probabilities via simulation for non-parametric, semi-parametric, and parametric duration models in Stata and R. Further, we also discuss how our interpretation strategy has useful implications for binary time-series-cross-section analyses. We use applications from judicial politics and interstate conflict to demonstrate our points.

---

[*] The authors' names appear in reverse alphabetical order. Paper prepared for presentation at the International Methods Colloquium (IMC). Draft, please do not circulate without the authors' permission.

What factors increase or decrease the probability that an event occurs? The question is a familiar one in political science. Wars, votes, reelection, policy adoption, democratic reversals, and treaty formation are some of the many events whose occurrence we have an interest in understanding. Yet for many of the events we study, it is reasonable to suspect that a subject's likelihood of experiencing the event is, at least in part, dependent upon how long the subject has been at risk. This suggests that events exhibit duration dependence: the probability of an event occurring in time period j *depends on* t, the amount of time elapsed since the subject could first experience the event. The duration dependence also suggests that for any question where we ask whether an event occurs, we can also ask *when* it occurs.

Duration models are a natural choice for modeling questions about when an event happens (Box-Steffensmeier and Jones 2004). They allow researchers to model the likelihood of an event occurring with a built-in ability to account for duration dependence, making it easy to evaluate when-related claims. However, no best practice exists for substantively interpreting duration model output. Many practitioners simply stop after interpreting the model's estimated coefficients, even though the estimates themselves are not directly interpretable. When substantive interpretation does happen, practitioners use a variety of strategies, such as expected durations, median durations, marginal effects, first differences, hazard ratios/rates, survival curves, cumulative hazards, or percentage changes in the hazard rate. Interpreting the hazard rate is among the more prevalent methods, and yet, hazards are not intuitive quantities with an easily interpretable meaning for many. As a consequence, many scholars tend to eschew duration models, opting for other models with both (a) more interpretable post-estimation quantities and (b) the capability to correct for duration dependence.[1]

We introduce a new, easy-to-understand quantity for interpreting duration models: transition probabilities. Transition probabilities are akin to logit/probit's predicted probabilities, and can be generated from non-parametric, semi-parametric, or parametric duration models. They permit researchers

[1] One prominent example is a logit/probit model with a covariate (or set of covariates) to account for duration dependence (e.g., splines, time polynomials [Beck, Katz, and Tucker 1998; Carter and Signorino 2010]). Scholars then generate predicted probabilities to interpret the logit/probit's results.

to make statements about the probability of a subject failing by time *t*, given some set of starting conditions (usually, that the subject just became at risk—$t = 0$). Conceptually, transition probabilities divide a process of interest into discrete stages, and provide the probability that a subject will be in one stage or any number of others. For instance, a country may be either democratic or autocratic, and transition probabilities provide an estimate of the likelihood a country occupies each of those stages.

We advocate transition probabilities' use because probabilities are far more intuitive quantities for readers to engage with than the current alternatives in the duration model literature. This is true particularly for hazards, which we repurpose and use to calculate transition probabilities. Confidence intervals are also easy to calculate, as we compute transition probabilities via simulation. Ultimately, transition probabilities help researchers to make their duration model's results more accessible to readers (King, Tomz, and Wittenberg 2000). We show how transition probabilities can simulated using an existing package in `R` (`mstate`) for non-parametric and semi-parametric models, or using a new package of our own creation in Stata for the same two model variants.[2] Our interpretation strategy should be of interest to frequent duration model users.

Our second claim is a powerful implication of our first. We focus on semi-parametric duration models in particular. We argue that researchers working with binary time-series-cross-sectional (BTSCS) data should consider using semi-parametric duration models over logit/probit in certain situations. We offer three principal reasons. First, semi-parametric models eliminate possible bias from misspecifying the event's baseline hazard, which denotes the underlying risk of the event's occurrence over time. Misspecification can arise in logit/probit because L/P's duration dependence correction (e.g., splines) explicitly parameterizes the baseline hazard, giving it a functional form. Misspecifying the baseline hazard's functional form can lead to inefficient estimates at best, and biased estimates at worst (Box-Steffensmeier and Jones 2004, 21–22). By contrast, semi-parametric duration models make no

[2] For a Stata package capable of estimating transition probabilities from parametric models, see Crowther and Lambert (forthcoming).

assumption about the baseline hazard's functional form, eliminating one possible source of specification bias.

Second, semi-parametric duration models force scholars to be more cognizant of the proportional hazards (PH) assumption. PH says any covariate's effect on the risk of an event occurring is proportional over time. If the PH assumption is met, for an event five times more likely to happen when $x = 1$ vs. $x = 0$ at time 1, the event will also be five times more likely to happen when $x = 1$ vs. $x = 0$ at every other time point. PH is discussed extensively and often in the semi-parametric duration model literature, particularly how to detect PH violations and correct for them. Yet, logit models also assume covariate effects are constant over time—that is, they make a PH assumption, too.[3] Importantly, this suggests that, if left unmodeled, BTSCS logit and probit models may suffer from misspecification and omitted variable bias, resulting in biased coefficient estimates.

Finally, semi-parametric duration models are a powerful gateway to modeling more complex processes. Political processes are often characterized by considerable complexity in the timing and order in which events occur, presenting difficulties for the structure of datasets and appropriate methods of analysis. McGrath (2015) outlines a common example: many studies' datasets fail to differentiate between the *onset* of an event and the *duration* of an event, either simply coding both as 1 or coding ongoing years as 0. Either of these potential solutions presents challenges to inference, and can mask important substantive results stemming from different covariate effects on each distinct outcome. In the context of logit/probit models, adequately modeling onset vs. duration while simultaneously accounting for different covariate effects is quite laborious. Semi-parametric duration models, in contrast, can readily accommodate this and more complex situations.

The paper proceeds in five parts. First, we review current interpretation strategies for duration models, pointing out their strengths and weaknesses. Second, we elucidate our proposed interpretation technique, transition probabilities—specifically, how they derive from a duration modeling framework.

[3] Carter and Signorino (2010) make the same point.

Third, we discuss how using transition probabilities in BTSCS analyses can improve our substantive interpretation of duration models, while also avoiding some of the pitfalls associated with logit/probit BTSCS models. Fourth, we provide two example applications, both of which were originally estimated using logit models with BTSCS data. Our fifth and final section concludes.

I. Current Interpretation Techniques

To start, we assume the reader has basic familiarity with duration models, which are concerned with lengths of time as dependent variable.[4] From that familiarity, we highlight two things that can affect which current interpretation strategies are applicable. First, duration models can be non-parametric, semi-parametric, or parametric. Second, a duration model's estimates may be reported in one of two metrics: accelerated failure time (speaking to t's length) or proportional hazards (speaking to t's terminating event).

All duration models are non-linear in parameters. As such, we cannot directly interpret the magnitude of any β's in our model. We are forced to generate additional quantities to present our model's results concretely (King, Tomz, and Wittenberg 2000). We loosely group extant approaches to interpretation along two major dimensions, each with two categories: the quantity of interest, and whether the quantity is an absolute level or a difference of some kind. The result is four different groupings:

TABLE 1. Current Approaches to Interpretation

Quantity	Absolute (Levels)	Relative (Difference)
t	Mean or median duration [$E(t)$ or $Q_{50}(t)$]	Marginal effect/ first difference [$\partial t/\partial x$, $\Delta E(t)$ or $\Delta Q_{50}(t)$]
hazard	Hazard rate [$h(t)$]	Hazard ratio/ % change in hazard rate [$\exp(\beta_{PH})$, $\%\Delta h(t)$]

[4] For a refresher, see our own Appendix B for a brief overview, and Box-Steffensmeier and Jones (2004) for a more thorough treatment.

In general, it is good practice to generate quantities that align with the research question/hypothesis' framing, as others have discussed at length. For instances where duration models are appropriate, two possibilities exist. If the stated research question is about durations—e.g., how long do democratic regimes endure—then quantities related to durations are more appropriate. If the research question is about the duration-terminating event—under what conditions do democratic regimes backslide—hazards are more appropriate. Methodologically speaking, there is no right or wrong answer. We bracket this reason in our subsequent discussion, to more squarely place the focus on what each interpretation technique does well or poorly.

Our theme throughout is that predicted durations are far easier to interpret than predicted hazards. However, some duration models cannot generate predicted durations easily, making hazard-based interpretations the only option. Hazards have few, if any, quantities that are easy to interpret. Our proposed interpretation technique fills this specific need by using hazards to generate transition probabilities.

1. Expected Duration/Median Duration

One way to interpret duration models is to predict *t* from its estimates, akin to how we can predict *y* from OLS models. In OLS, we take expectations to yield:

$$E(y|X) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

which is what we usually call "predicted *y*" ($\hat{y}$).

Duration models are more complicated, because some of the parametric models' underlying distributions are not symmetric. The Weibull parametric duration model is one example. Because of the non-symmetry, two quantities may be of interest to us: the mean duration time (via expectations) and the median duration time. The expression for each quantity varies depending on the baseline hazard's functional form (for formulas, see Box-Steffensmeier and Jones 2004, chap. 3).

This interpretation technique works well for any parametric duration model, regardless of metric. Deriving the mean and median *t* for semi-parametric Cox models is more difficult, because both formulae

involve an assumption about the baseline hazard's functional form. The Cox's defining feature is that it does *not* assume anything about the baseline hazard's functional form. Kropko and Harden (2015) provide a technique for recovering expected durations from a Cox model, by using a generalized additive model (GAM) to regress the Cox's predicted exp $(\beta_{PH}{}'X)$ against t, the observed duration. For non-parametric models, the mean and median are also derivable expressions by way the Kaplan-Meier estimate of $S(t)$ (see Klein and Moeschberger 2003, 32–33).

For mean or median durations to be useful, however, the quantities must be calculated for different x values, since our hypotheses pertain to how t responds to changes in x. Plotting the mean/median duration graphically for different x values, along with each quantity's corresponding confidence interval, allows us to say whether the mean/median duration for one x value is statistically distinguishable from the mean/median duration for another x value, based on whether the confidence intervals overlap.[5] Graphing the quantities of interest and checking confidence interval overlap is arguably the most popular way of presenting and interpreting predicted probabilities from logit/probit.

2. Marginal Effects/First Differences

We can also calculate the effect of a change in x on the expected duration, if we so desire.[6] It is akin to interpreting β in OLS, because OLS β estimates are equivalent to marginal effects. If x is continuous, we can take the partial derivative of $E(t|X)$ with respect to x, yielding x's marginal effect. Marginal effects tell us how much $E(t|X)$ changes for an infinitesimally small increase in x's value from a (0, or 1, 2…). If x is non-continuous, we can take the discrete first-difference. The first difference would tell us how much $E(t|X)$ changes for a one-unit increase in x from some value. First differences are also perfectly acceptable for a continuous x. Licht (2011, 230) nicely summarizes the differences between

[5] For the dangers of making multiple comparisons like this, and how to adjust for these dangers, see Esarey and Sumner (2016).

[6] If we wanted, we could also calculate x's marginal effect on the median duration, instead of $E(t|X)$ (Box-Steffensmeier and Jones 2004, 31).

marginal effects and first differences, in the context of interpreting interaction effects in duration models. Her summary generalizes to models without interaction terms.

In practice, we prefer calculating first differences, regardless of whether x is continuous or not. We find it easier to speak of specific changes in x's value, instead of a nebulous "infinitesimally small increase," and how those specific changes correspond to changes in the mean duration. Further, when we look at graphs of predicted quantities from different covariate profiles (like the previous subsection) and check for overlapping confidence intervals, we are in fact visually checking the discrete differences (first or otherwise), and seeing if the differences are statistically distinguishable from zero. However, we reiterate that both marginal effects and first differences are correct, from a technical perspective.

3. HAZARD RATE

In continuous-time duration models, the hazard represents the instantaneous rate at which the termination event occurs. We can interpret x's effect with respect to the hazard of the event occurring. Hazard-related interpretations are common for duration models whose βs come only in a PH metric, with Cox models being the canonical example. The generic form for any hazard is:

$$h(t) = h_0(t)\exp\left(\beta_{PH}{}'X\right) \qquad 1$$

The first hazard-related interpretation involves predicting $h(t)$'s value for a particular covariate profile, similar to how we can predict t for some models. $h_0(t)$'s functional form will vary, depending on what type of duration model is estimated. Our earlier remarks about graphing and confidence intervals in the context of mean/median durations also apply here.

This interpretation technique works well for parametric duration models, because we impose a functional form for $h_0(t)$. The technique works less well for Cox models, since we do not impose a functional form for $h_0(t)$, but it is possible to recover an estimate for $h_0(t)$ after estimating the model. Some software packages provide a canned procedure for plotting a smoothed estimate of a Cox's $h(t)$, which automatically recovers $h_0(t)$ when creating the plot. In lieu of a canned procedure, the cumulative

hazard, $H(t)$, is an easier quantity to plot for Cox models because it can be estimated from the survivor function: $H(t) = -\ln(S(t))$.

The major drawback for hazard rates is their interpretability. Hazards, as a predicted quantity, are arguably unique to duration models with little analogue elsewhere.[7] At best, they represent a conditional probability (discrete-time durations): the probability that the duration's terminating event occurs, given that it has not occurred yet. At worst, they represent an instantaneous risk (continuous-time durations), with a similar interpretation as marginal effects: it is the risk of experiencing the duration's terminating event for an infinitesimally small increase in t's value. Neither possibility is particularly appealing.

3.5. Cumulative Hazard and Survivor

Researchers also use two other absolute quantities to interpret duration models, which are particularly prevalent techniques for interpreting semi-parametric models. The first is $H(t)$, the cumulative hazard, which we mentioned above. $H(t)$'s slope speaks to the rate at which $h(t)$ changes, and can be larger than 1. $H(t)$ can be useful for comparing various covariate profiles, because $H(t)$ graphs will show which covariate profile has the highest risk of the event occurring to date. However, $H(t)$'s biggest drawback is that its magnitude is not informative on its own (Singer and Willett 2003, 488). Researchers require additional context to get a sense of an event's probability of occurrence *overall*. For instance, a cumulative hazard equal to 1.2 for a particular covariate profile tells us nothing. We would need another covariate profile or two, to be able to get a sense of whether this value is large or small by making relative comparisons.

The second quantity is $S(t)$, the survivor function, defined formally as $S(t) = \Pr(T \geq t)$. The survivor function provides information about how many subjects' failure times (denoted T in the formula) are above a specific time (usually denoted t). For instance, if we had five subjects who each failed at $t =$

[7] The notable exception is sample selection models. The selection equation generates the inverse of Mills' ratio (IMR), which is then included as a regressor in the outcome equation. The IMR is, in fact, a hazard (Heckman 1979).

1, 2, 7, 8, and 10, $S(t = 5) = 0.6$, since three subjects have failure times larger than 5. We do like the survivor function, because it is a close relative of transition probabilities, the technique we advocate. There are highly specific situations where our transition probabilities will actually equal $F(t) = 1 - S(t)$.[8] However, $S(t)$ (or $F(t)$) is less useful for interpreting more complex duration models, such as competing risks and repeated events. Transition probabilities give practitioners more flexibility to interpret these and other possible simulation scenarios (for example, seeing how long until a subject fails if it was still alive at $t = 10$), generalize to all possible transition structures (e.g., competing risks, repeated events), and also generalize to all duration model types. In addition, we also rarely report confidence intervals around $S(t)$—or any other post-estimation duration model quantity, for that matter. Confidence intervals are simply not the current prevalent norm, and yet, our ultimate interest in making inferences suggests it should be. By contrast, we can easily obtain confidence intervals around transition probabilities, since we compute the probabilities via simulation.

4. Hazard Ratio, Percent Change in Hazard Rate

For relative comparisons using the hazard, one option is the hazard ratio, also sometimes called the risk ratio or relative risk (Box-Steffensmeier and Jones 2004, 50; Mills 2011, 94). The hazard ratio is equal to $\exp(\beta_{PH})$, and it describes "the effect of a one-unit difference in the associated predictor on the…hazard" (Singer and Willett 2003, 524). It is a ratio because it compares $h(t)$ from one covariate profile ($x_{(1)} = 0$, say) to $h(t)$ from another covariate profile ($x_{(2)} = 1$), and expresses how to obtain $h(t|x_{(2)})$ as a multiple of $h(t|x_{(1)})$. Or, more plainly: if covariate profile $x_{(1)}$'s hazard value represents an image, $\exp(\beta_{PH})$ represents how much you would have to shrink or enlarge that image to obtain covariate profile

[8] The two will be equal for a classic duration transition structure—all subjects start in stage 1, two possible stages, and the only transition possible is 1→2—when the transition probability interval begins at $s = 0$. In a similar vein, our transition probabilities will equal the cumulative incidence function (CIF) in a classic competing risks situation—all subjects begin in stage 1, there are any number of additional stages, and the only possible transitions are out of stage 1—again when the transition probability interval begins at $s = 0$ (Putter, Fiocco, and Geskus 2007, 2424).

$x_{(2)}$'s hazard. This is evident from inserting $x_{(1)} = 0$ and $x_{(2)} = 1$ as arbitrary values into equation 1 and simplifying terms:

$$h\left(t|x_{(1)}\right) = h_0(t)\exp(\beta_{PH}x_{(1)}) \qquad h\left(t|x_{(2)}\right) = h_0(t)\exp(\beta_{PH}x_{(2)}) \qquad 2$$

$$\frac{h\left(t|x_{(2)}\right)}{h\left(t|x_{(1)}\right)} = \frac{\exp\left(\beta_{PH}x_{(2)}\right)}{\exp\left(\beta_{PH}x_{(1)}\right)}$$

$$\frac{h(t|x=1)}{h(t|x=0)} = \exp(\beta_{PH}[1-0])$$

$$h(t|x=1) = \exp(\beta_{PH})h(t|x=0)$$

This interpretation technique is appealing for Cox models, because the baseline hazard terms cancel out in the ratio, eliminating the complications arising from $h_0(t)$ in a semi-parametric expression. Because of the exponentiation, we are interested in seeing whether $\exp(\beta_{PH})$ is statistically different from 1 for hypothesis testing purposes, since exp(0) = 1. A hazard ratio greater than 1 means that a $(x_{(2)} - x_{(1)})$-unit change from $x_{(1)}$ produces a higher hazard rate. A hazard ratio less than 1 means that a $(x_{(2)} - x_{(1)})$-unit change from $x_{(1)}$ produces a lower hazard rate. The hazard ratio's major disadvantage is yet again its interpretability. Hazard ratios are not particularly intuitive quantities, perhaps because they refer to the hazard's scale across two different covariate profiles. This is especially true for audiences not already familiar with hazards, as we have already mentioned.

Finally, we can compute x's effect on $h(t)$ via the marginal effect or first difference. The generalities of our marginal effects/first difference discussion, when t was the quantity of interest, also apply here. For $h(t)$, the percent change in the hazard's value is equivalent to the first difference. To calculate how some change in x's value affects $h(t)$ in terms of percent change, the formula is (Box-Steffensmeier and Jones 2004, 60; Mills 2011, 95):

$$\%\Delta h(t|X) = \frac{\left(\exp\left(\beta_{PH}x_{(2)}\right) - \exp\left(\beta_{PH}x_{(1)}\right)\right)}{\exp\left(\beta_{PH}x_{(1)}\right)} * 100 \qquad 3$$

$$\%\Delta h(t|X) = \left(\exp\left(\beta_{PH}\left[x_{(2)} - x_{(1)}\right]\right) - 1\right) * 100$$

where $x_{(1)}$ and $x_{(2)}$ denote the x values of interest. Earlier, we used $x_{(1)} = 1$ and $x_{(2)} = 0$.

Box-Steffensmeier and Jones (2004, 60) showcase $\%\Delta h(t|X)$ as a preferred way for interpreting Cox models, for logical reasons. First, Cox model estimates can only be expressed in PH, as we mentioned earlier. Percent changes are marginally more intuitive than hazard ratios, if only because the same type of percentage calculations appear with much greater frequency in many econometric and non-econometric contexts. In addition to the above, either knowingly or unknowingly, practitioners usually compute $\%\Delta h(t|X)$ to make meaningful sense of the hazard ratio for one-unit increases in x. For instance, if x's hazard ratio ($\exp(\beta_{PH})$) is equal to 0.6, it can be equivalently expressed as $0.6 = \exp(\beta_{PH}*[1 - 0])$, which is the entire first term in equation 3's formula. Plugging in 0.6 and finishing the calculation shows a one-unit increase in x produces a 40% reduction in the event's hazard, as $0.6 - 1 = -0.4$.

$\%\Delta h(t|X)$ does nevertheless have limitations, aside from $h(t)$'s interpretability. In particular, as with any method focusing on percent changes in y's value, the quantity may be misleading for relatively infrequent events. For instance, it may be that increasing x's value produces a 100% increase in the hazard rate. However, in substantive terms, this may be a very small change in the likelihood of an event actually occurring. A 100% increase the hazard could come about if $h(t)$ increased in value from 0.4 to 0.8 (fairly frequent event), but it could also come about if the hazard increased from 0.00001 to 0.00002 (very infrequent event). We simply cannot tell from percentage changes alone.

To succinctly summarize our discussion: predicted durations (mean or median) are typically easiest for readers to grasp with minimal effort. These can be generated from any parametric duration model. Semi-parametric models have difficulties generating predicted durations, because the formulae require $h_0(t)$ to be parameterized. Notably, this includes the Cox model, the workhorse of semi-parametric duration models. Predicted hazards are the only alternative, but hazards are not particularly intuitive quantities to understand and interpret, regardless of a practitioner's adeptness with duration models. Relative measures of comparison involving $h(t)$ exacerbate the situation, since their calculation involves taking an already hard-to-interpret quantity and transforming it further.

II. Proposed Technique: Transition Probabilities

A. Origin and Formulae

Our transition probabilities address the need for easier-to-understand interpretation techniques for hazard-based quantities, which has powerful implications for semi-parametric Cox models. The idea of obtaining *transition probabilities* from a duration model has its firmest roots in the multi-state duration model literature, but is generalizable to all duration models. Usually, multi-state duration models are Cox models estimated with unique baseline hazards and unique covariate effects for every event within a process (see Metzger and Jones 2016 for an overview).[9] However, multi-state duration models can also be estimated parametrically or non-parametrically, meaning transition probabilities can also be calculated using any of the three duration model variants.

Multi-state duration models begin by recognizing we can view any process as being composed of a number of stages. Each stage is defined based on the event(s) subjects are at risk of experiencing—"risk sets," in duration parlance. Simple duration models have two stages: (1) at risk of failing, which all subjects occupy to begin, and (2) failed, once subjects experience a failure event. A subject experiencing the event moves *from* one stage *to* another. We use the word "transitions" to denote this movement between stages, with its directed from-to pairings. The emphasis on stages opens up a new way of conceptualizing the output from duration models. Instead of thinking purely about *when* a subject experiences a transition, we can ask *which* stage a subject occupies at a given point in time. Answering a "which?" question requires information about *whether* a transition event occurred and *when* it occurred, but goes one step further by speaking to *which* risk set a subject likely occupies by t.

Defining transition probabilities begins with a familiar quantity: the hazard, often called a "transition intensity" in the multi-state literature and denoted $\alpha(t)$ instead of $h(t)$. We use a as a generic identifier for a subject's current stage ("from"), and b as the generic identifier for a subject's next potential stage ("to"). If we add transition-specific notation to the generic expression for a hazard, we

[9] If needed, we can constrain different combinations of baseline hazards and/or the covariates' coefficients to be equal.

obtain the instantaneous risk of transitioning from Stage *a* to Stage *b* at time *t* (Wreede, Fiocco, and Putter 2010, 262):

$$h_{a\to b}(t) \equiv \alpha_{a\to b}(t) = \lim_{\Delta t\to 0} \frac{\Pr\left(Z(t+\Delta t) = b | Z(t) = a\right)}{\Delta t} \qquad 4$$

where $Z(t)$ denotes the random process determining the stage's value in t (Stage a, Stage b, etc.).

Our earlier discussion of hazards' strengths and weaknesses, in addition to what we mentioned in the last paragraph, makes clear that hazards and probabilities are similar, but not synonyms. The challenge becomes figuring out whether, and how, one of the quantities can be expressed in terms of the other, in such a way to yield transition probabilities. Gill and Johansen (1990, 1532–1534) make the key connection in their work on product integrals. They recognize that the cumulative hazard function[10] and the survivor function belong to a family of interval functions with well-known properties, drawing on the counting process literature. Gill and Johansen use these properties to map the cumulative hazard onto the survivor, and subsequently show that this relation generalizes to more complex settings, where we have a process composed of many transitions, not just one (see also Aalen, Borgan, and Gjessing 2008, Appendix A; Andersen et al. 1993, 88–95). As a consequence, we can aggregate every transition's cumulative hazard into an S x S matrix, $\mathbf{H}(t)$ (in the multi-state literature, $\mathbf{A}(t)$), where S is the number of stages within the process. After forming $\mathbf{H}(t)$ and applying Gill and Johansen's basic results, we obtain:

$$P(s,t) = \prod_{u\in(s,t]} (\mathbf{I} + \Delta\mathbf{H}(u)) \qquad 5$$

where u denotes all times at which we observe any transition within some time interval with start point s and end point t. The resultant quantities are now best described as transition probabilities from a Markovian process, contained in the $P(s,t)$ matrix. The matrix's quantities represent the probability of transitioning from each stage to every other stage within the time interval[11] $(s,t]$. For example, element

[10] The cumulative hazard function, $H(t)$, simply sums the hazard's value at every time point from 0 through t. In a non-parametric setting, it is equivalent to the Nelson-Aalen estimator.

[11] The "within the time interval" in the interpretation stems from the order in which each $\mathbf{I} + \Delta\mathbf{H}(u)$ term is multiplied together by the product integral.

$P_{2,1}(s,t)$ would denote the probability of a subject transitioning from Stage 2 to Stage 1within the time period between s and t. Metzger and Jones (2016, Appendix A) walk through a transition probability matrix calculation using a simple competing risks process.

As we mentioned above, equation 5 is contingent on the process having transition-specific hazards that are Markovian: a subject's next stage is conditional only on the subject's current stage. Whether this is true can be checked via transition-specific hazards. Markovian transition-specific hazards have values that do "not depend on any other aspect of the history, like states visited on the way, and the times of previous transitions (except to the extent that this information is reflected by the present state)" (Hougaard 2000, 143). If there are covariates that capture any of this information, the model is semi-Markovian in nature. If durations are recorded as gap time, the model is also semi-Markovian, since t is resetting *any* time a transition occurs, implying that t tells us something about the transition history. Transition probabilities must be simulated for semi-Markov models, instead of derived analytically.

B. Generating in R/Stata

We elect to compute all our transition probabilities via simulation, for a few reasons. First, simulations can handle non-, semi-, and parametric duration models. Analytic expressions tend to be more rigid in their formulation, specific to a particular distribution or broader class of models. Second, our simulations do not make a Markov assumption about stages to estimate, whereas analytic expressions do. Third, our simulation setup is flexible and can accommodate a variety of risk set configurations. This includes, but is not limited to, situations with recursive, repeated, competing, and/or sequential events. Analytic expressions have difficulty handling some of these alternatives, particularly recursive situations in which subjects can experience an event, and then once they experience a second event, become at risk of experiencing the first again. To get the transition probability analytically, researchers must be able to express every possible transition sequence a subject can take—a difficult proposition with recursiveness. Finally, a simulation approach makes generating confidence intervals quite simple, whereas analytic derivations are often quite complex.

Wreede, Fiocco, and Putter's (2011) `mstate` package in `R` can calculate both analytic and simulated transition probabilities. The simulations are configured as a series of nested risk experiments. Our Stata package uses simulations exclusively, following the same setup as Wreede, Fiocco, and Putter (2011), which itself makes use of Dabrowska (1995). Beyersmann, Allignol, and Schumacher (2011) also use the same setup to structure their book.[12]

III. Implications: Modeling with BTSCS Data

A. Current Approaches to BTSCS

Beck, Katz, and Tucker's (1998, hereafter "BKT") article on BTSCS data lays out the situation well. A sizable swath of political science research is concerned with whether or not an event occurs, making the dependent variable binary. We have time-series cross-sectional data on multiple panels, with each subject *i*'s panel composed of observations from multiple points in time (*j*, to denote calendar year/time). Data of this form are binary time-series cross-sectional data (BTSCS). They are a form of grouped duration data. Logit and probit (hereafter, L/P) models are the current standard tools for evaluating hypotheses about an event's occurrence. We typically generate predicted probabilities to interpret our L/P model results.

For many of the events we study, it is reasonable to suspect that temporal dependence exists: whether the subject experiences the event in the current time period *j* is dependent upon how long the subject has been at risk of experiencing the event. This length of time is *t*, our duration.[13] However, by default, L/P both assume temporal *in*dependence.[14] No covariate captures *t*'s effect on the probability of the event occurring in *j*. BKT (1998) point out the importance of accounting for temporal dependence because failing to do so amounts to assuming an incorrect functional form for the baseline hazard. A

[12] We give a broad outline of the procedure in Appendix C.

[13] *t* is more properly expressed as t_{ij} in a BTSCS context, as the variable's value is specific to a subject (*i*) at one point in (calendar) time (*j*). We omit the subscripts to streamline our presentation.

[14] The exponential parametric duration model is the equivalent for continuous-time data, as both the exponential and L/P assume a flat baseline hazard.

wrong $h_0(t)$ functional form in BTSCS has the same ramifications as it does for parametric duration models: inefficient estimates at best, biased estimates at worst (BKT 1998, 1261; Box-Steffensmeier and Jones 2004, 21).

BKT suggest correcting for temporal dependence by including either t dummies[15] or some function of time (e.g., splines) as regressors. Time dummies are plagued by estimate inefficiency and potential separation problems when estimated parametrically (Carter and Signorino 2010, 275), making some function of time the more appealing option for many applications. Cubic polynomials (t, t^2, and t^3) are the current (albeit contested, by some (Beck 2010)) incumbent correction, because of how easy they are to generate and their flexibility in capturing many different forms of temporal dependence (Carter and Signorino 2010). BKT (1998, Appendix A) show that a continuous-time Cox model is identical to a BTSCS complementary log-log (cloglog) model with time dummies.[17] The βs returned by both models are mathematically equivalent. The larger implication of this is that Cox models can be used in lieu of L/P for modeling BTSCS data, since cloglog and L/P are both binary outcome models. Computing transition probabilities from our Cox model gives us familiar predicted quantities in the BTSCS setting, eliminating "difficult interpretation" as a reason for choosing L/P over Cox.

B. Limitations of Logit/Probit for Modeling BTSCS Data

At minimum, there are three possible advantages to using a Cox model for analyzing BTSCS data. First, and most important in the majority of applied work, is the assumption regarding proportional hazards. While this assumption is most closely associated with the Cox model, it nevertheless is an assumption made by logit/probit models for BTSCS data. Consequently, much as is the case with Cox models, it is necessary to test for violations of this assumption in logit/probit models for BTSCS data, a point that Carter and Signorino (2010, 289) readily make.

[15] Time dummies are dichotomous variables for every t we observe. The dummies serve as fixed effects, permitting each t to have a unique intercept—i.e., baseline hazard.

[17] We show the derivation in Appendix A.

However, in practice, testing for violations of this assumption in a logit/probit framework is cumbersome. The test requires estimating a series of likelihood ratio tests involving the core restricted model and an unrestricted model that includes an interaction between the time polynomials and a covariate, for every covariate in the model (Carter and Signorino 2010, 289). As a result, in many applied settings, practitioners will not follow Carter and Signorino's (2010) advice and will not test for PH violations in a logit model. The result of a PH violation in logit is the same as it would be in a Cox model—misspecification and potentially erroneous, or limited inferences. By contrast, testing for PH violations after a Cox model is trivial, as canned routines exist in standard software packages employed by political scientists.

Second, using semi-parametric models removes a possible source of bias stemming from misspecifying the baseline hazard. Misspecification can arise in L/P because in order to model the baseline hazard, L/P explicitly parameterize the baseline hazard, often using a cubic polynomial.[18] Such misspecification can potentially lead to either inefficient or biased estimates parameter estimates (Box-Steffensmeier and Jones 2004, 21–22). Moreover, how the baseline hazard is modeled in L/P, can have a significant impact on the subsequent ability to properly interpret the hazard. In contrast, the use of a Cox model avoids these issues altogether.

Finally, Cox models can easily handle causally complex processes. Jones and Branton (2005) emphasize the similarity between BTSCS setups and Cox models, noting the Cox model's ability to handle additional process structures that L/P cannot—namely, competing and repeated events. Metzger and Jones (2016) extend Jones and Branton's list, showing that Cox models can also handle additional process structures when formulated as a multi-state model, such as recursiveness and sequential events. Metzger and Jones (2016) subsequently show the importance of modeling all the process' stages, to protect against inaccurate and misleading inferences about *x*'s effect. In short, using Cox models in a

[18] Automated smoothing splines avoid some of these issues, though in practice they may prove challenging to implement given the range of options available to adjust their fit, and the risk of overfitting as a result.

BTSCS setting can actually give researchers *more* modeling freedom than L/P currently offers, giving us more leverage to test additional implications from our theories.

IV. Applications

We employ two applications below to demonstrate and explore the use of transition probabilities in comparison to other means of interpretation. In the first, we re-examine Hall and Ura's (2015) study of judicial invalidation of major federal laws to demonstrate the advantages of transition probabilities in comparison to other post-estimation techniques. In the second, we re-estimate the ubiquitous model of militarized interstate dispute onset (MID) (Oneal and Russett 2005), to show the additional benefits of employing Cox models vs. a more standard logit for PH violation tests and modeling causal complexity.

A. Judicial Review

We use Hall and Ura's (2015) study of US judicial review to demonstrate transition probabilities' use as an alternate means of substantively interpreting Cox models. Hall and Ura argue the judiciary is more likely to invalidate significant legislation if the legislature demonstrates little support for the legislation. To test this claim, Hall and Ura construct a dataset of statutes from 1949-2008. The unit of analysis is the statute-year, and the dependent variable is a dichotomous indicator of whether the Supreme Court invalidates the statute in a particular year. Hall and Ura employ three different independent variables, each of which attempts to measure the degree of support for the statute by the current pivotal legislator. Each variable captures the likelihood that the pivotal legislative actor would vote in favor of the law, but varies who is defined as being pivotal. The Floor Median Model focuses on each chamber's median member as well as the president, the Senate Filibuster Model includes filibuster actors, and the Party Gatekeeping Model includes the majority party's median member in both the House and the Senate (2015, 823).

To model the relationship between legislative support for a statute and the likelihood of judicial invalidation, Hall and Ura estimate a series of logit models. As part of their model specification, they

include a counter of the number of years since a statute was passed, along with cubic polynomials, as suggested by Carter and Signorino (2010). In explaining their modeling strategy, Hall and Ura state, *correctly* we would argue, that their "approach is functionally equivalent to a traditional duration analysis *and offers clearer interpretation*" (2015, 824, emphasis added).

1. Model Estimates and Conventional Interpretation

Table 2 replicates all three of Hall and Ura's logit models, and re-estimates each as a Cox model. As the coefficient estimates in Table 2 indicate, both the logit and Cox estimation strategies yield similar inferences. Regardless of how legislative support is operationalized, the higher legislative support for a particular statute, the less likely it is that the Supreme Court will overturn the statute. However, focusing solely on the coefficient estimates' sign limits researchers' ability to substantively interpret their results. In the main text of their manuscript, Hall and Ura use predicted probabilities and marginal effects to substantively interpret their logit models' estimates.

Table 2. Comparison of Logit and Cox Estimates of Judicial Review

	Floor Median		Senate Filibuster		Party Gatekeeping	
	Logit	Cox	Logit	Cox	Logit	Cox
Majority Support	-1.50**	-1.47*	-1.14*	-1.11†	-1.32**	-1.28*
	(0.51)	(0.66)	(0.50)	(0.62)	(0.49)	(0.59)
Constant	-2.59**	--	-2.97**	--	-2.88**	--
	(0.51)		(0.45)		(0.43)	
Log-Likelihood	-272.5	-264.6	-273.7	-265.8	-272.8	-264.9

† = $p \leq 0.10$, * = $p \leq 0.05$, ** = $p \leq 0.01$, two-tailed tests. Partial log-likelihood estimates reported for Cox models. Logit models also include a counter for years since a statute passed, along with quadratic and cubic terms.

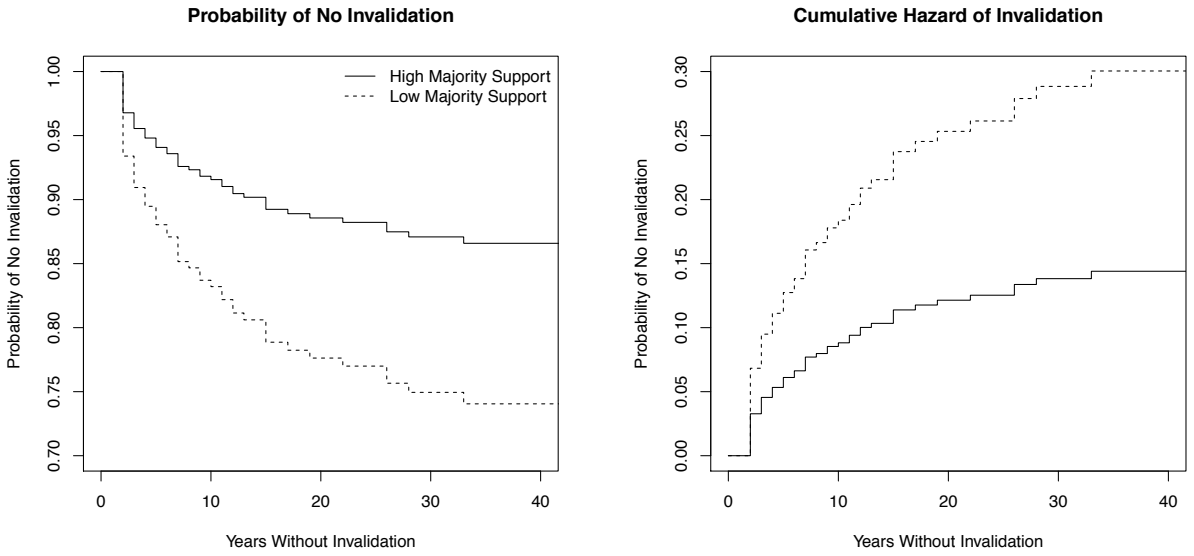
We begin by interpreting Table 2's Cox model using the current interpretation strategies we discussed earlier. Figure 1(a) compares two survival curves, one depicting the probability a statute will stand with high legislative support in the Floor Median Model (solid line), and the other depicting the probability it will stand with low legislative support in the Floor Median Model (dotted line).[19]

[19] We focus on the Floor Median Model results, to facilitate a comparison with Hall and Ura's estimated predicted probabilities.

Consistent with the Hall and Ura's estimated predicted probabilities, Figure 1(a) indicates that statutes with high legislative support are less likely to be invalidated by the Supreme Court than statutes with low levels of legislative support. Moreover, regardless of the level of legislative support, statutes become more likely to be invalidated over time, though this effect slows considerably. However, our two drawbacks to survival curves (or $S(t)$'s complement, $F(t)$) are also evident: we can speak only of what happens immediately after a law is passed (starting time = 0); and reporting $S(t)$'s confidence intervals is not the current norm, in part because they are hard to calculate in some statistical packages.

FIGURE 1. Survival Curves and Cumulative Hazard Estimates of Judicial Invalidation

(a) Survival Curves (b) Cumulative Hazards



NOTE: High majority support is equal to 1 SD above the mean, and low support is equal to 1 SD below.

Figure 1(b) presents cumulative hazard estimates from the same model, using the same covariate profiles as Figure 1(a). Again, the results' substantive interpretation leads to the same conclusion as Figure 1(a). Laws with low majority support have a higher risk of being invalidated than laws with a high level of support. However, interpreting cumulative hazard curves also has challenges, especially compared to predicted probabilities. To echo our earlier discussion, notice first how Figure 1(b)'s y-axis values are only useful when comparing various covariate profiles, because we can see low majority support has the highest cumulative risk of occurring to date. However, also notice that interpreting $H(t)$

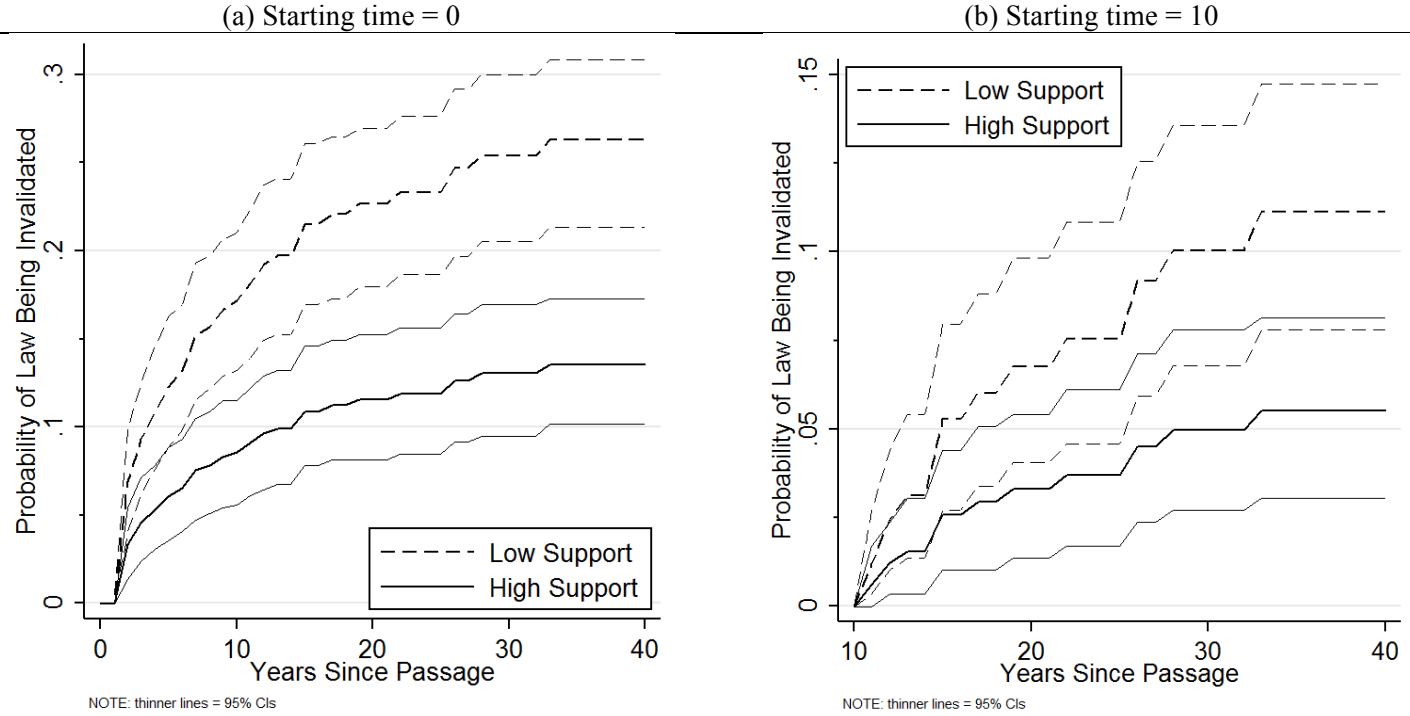
for a single covariate profile means relatively little. Is a cumulative hazard value of 0.30, for instance, a high value, a low value, or something in between?

Finally, we can use Equation 3 to estimate percent changes in the hazard rate, given change in a covariate's value. Increasing majority support's value from the mean to one standard deviation above produces a 30% reduction in the hazard that a law will be invalidated. Conversely, a 1-standard deviation reduction from the mean yields a 44% increase in the hazard that a law will be invalidated. Percent changes are useful for providing a basic picture of how changes in majority support impact the risk of a law being invalidated, but they provide little sense of the overall risk of invalidation. That is, a 44% increase may be meaningful in some instances, but not in others, depending on the underlying likelihood of invalidation.

2. Interpretation via Transition Probabilities

We argue that estimating transition probabilities from duration models provides more intuitive substantive interpretations than most current post-estimation strategies. To highlight these features, we estimate transition probabilities based on the same model of judicial review discussed above. Specifically, we simulate transition probabilities for two situations. In the first, we specify that the starting time is 0 and the starting stage is stage 1, meaning the law has just passed and has not yet been invalidated. In the second, we keep the same starting stage, but specify the starting time as 10, meaning the law was passed 10 years ago and has not yet been invalidated. The current debate about invalidating the Affordable Care Act underscores why a scenario with a non-zero starting time might be useful (the ACA was enacted in 2010). As above, we estimate these transition probabilities for two scenarios: one in which legislative support for the law is high and one where legislative support is low.

FIGURE 2. Transition Probabilities of Judicial Review



NOTE: Transition probabilities generated with 1000 simulations, $n = 295$. Starting stage = law valid (both panels)

Figure 2 plots the simulated transition probabilities. There are two possible stages in which a law may be situated: not yet invalidated and already invalidated. Importantly, the latter stage is absorbing—once a law has been invalidated, it is no longer at risk of being invalidated, nor is it at risk of being re-instated.[20] As a consequence, Figure 2's values reflect the probability that a law *has been* declared invalid by that particular year. For instance, the transition probability in $t = 5$, reflects the probability that the law was declared invalid in year 5, but also the probability that the law was declared invalid in $t = 4$, $t = 3$… $t = 1$. Notice how the probabilities reflect the overall likelihood that a law has been declared invalid in each year, regardless of *when* it was declared invalid, which is different from the probability of invalidation happening in any given year.

Figure 2 depicts several interesting results that demonstrate transition probabilities' utility. First, focusing on Figure 2(a), there is a relatively high chance that significant legislation will be invalidated.

[20] There is no requirement that such stages be absorbing. The second application provides an example without an absorbing stage.

Regardless of the level of legislative support, there is roughly a 20% chance that a major piece of legislation will be ruled invalid within 5 years of its passage, and a roughly 40% chance of such a ruling within 20 years of its passage. Starting with the 5th year after passage, laws with low legislative support have a significantly higher probability of being invalidated than laws with high legislative support, indicated by the non-overlapping confidence intervals. Focusing only on the probability of invalidation happening in any given year—which may be relatively low—risks obscuring these larger points about invalidation across time. Notice also how the figure's confidence intervals make it easy to evaluate whether the two scenarios' transition probabilities are statistically different from one another.

Second, and related to the prior point, transition probabilities are in a readily interpretable and intuitive scale. This facilitates comparisons between different values (e.g., is the likelihood of invalidation higher when legislative support goes down?), but also absolute inferences. A 50% chance that major legislation will be invalidated within 30 years of its passage clearly indicates preserving major legislation is far from guaranteed.

Third and finally, transition probabilities give us more flexibility to specify additional situations of interest. Figure 2(b) shows that, if a law has been on the books for 10 years, it will have a 6.8% chance of being invalidated in the presence of low legislative support at the 20-year mark, whereas it will have a 3.3% chance of being invalidated in the presence of high legislative support at the 20-year mark. However, these differences are not statistically distinguishable from each other, suggesting that *if* laws survive to the 10-year mark, the degree of legislative support has no significant effect on the law being invalidated in subsequent years. These sort of inferences, where our starting time for the simulations is not 0, are not possible with our current interpretation strategies.

B. MIDs

Our second application regarding MID onset demonstrates the potential advantages of using Cox models to deal with problems relating to duration dependence. In particular, Cox models more readily facilitate testing important statistical assumptions about proportional hazards (PH) and modeling causally

complex processes. Our data come from Oneal and Russett (2005), covering 1885-2000 with a dyad-year unit of analysis. In their dataset, the dependent variable is coded 1 for dyad-years in which the two states begin a new MID in that year and 0 otherwise. We end up recoding this dependent variable, for important reasons we explain below. Our dependent variable is coded 1 for any dyad-year in which two states have a new *or* ongoing MID and 0 otherwise. We use the classic "liberal-realist" specification in our example, which are a set of control variables fairly standard across studies of MID onset.

1. Model Estimates and PH Testing

Table 3's first two columns compare the logit and Cox model estimates for MID onset. To ensure these models are equivalent, we include a variable counting the number of years since the last MID onset, along with cubic polynomials, in the logit model. The two models' results are largely similar, and support most of the established conclusions regarding MID onset determinants. For instance, both models indicate higher levels of economic interdependence reduces the risk of dyadic MID onset, as do greater distances between the two states and larger disparities in the two states' relative capabilities. Conversely, being neighbors and having a major power in the dyad both increase the risk of dyadic MID onset.

TABLE 3. Comparison of Logit and Cox Estimates of MID Onset

	Logit	Cox	Cox with PH Corrections
Allies?	-0.205*	-0.081	-0.261**
	(0.090)	(0.062)	(0.091)
Capability Ratio (ln)	-0.165 **	-0.200**	-0.249**
	(0.031)	(0.018)	(0.026)
Economic Interdependence (Low)	-18.69**	-42.03**	-56.04**
	(7.159)	(5.69)	(8.768)
Contiguity	2.220**	1.965**	1.897**
	(0.131)	(0.066)	(0.067)
Distance (ln)	-0.841**	-0.415**	-0.325**
	(0.059)	(0.023)	(0.030)
Major Power Dyad	2.692**	1.984**	1.786**
	(0.113)	(0.057)	(0.081)
Democracy (Low)	-0.064**	-0.065**	-0.056**
	(0.007)	(0.005)	(0.007)
Joint IGOs	0.011**	0.021**	0.037**
	(0.002)	(0.002)	(0.003)
Allies * ln(Time)			0.103*
			(0.044)
Capability Ratio * ln(Time)			0.028*
			(0.014)
Econ * ln(Time)			10.79**
			(3.538)
Distance * ln(Time)			-0.082**
			(0.016)
Major Power Dyad * ln(Time)			0.130**
			(0.041)
Democracy * ln(Time)			-0.006†
			0.0034
Joint IGOs * ln(Time)			-0.012**
			(0.001)
Log-Likelihood (partial)	-7851.9	-15779.7	-15708.6

† = $p \leq 0.10$, * = $p \leq 0.05$, ** = $p \leq 0.01$, two-tailed tests. Logit model also includes peace years and cubic polynomials.

While the estimates from the two models are similar, the Cox specification presents a central advantage over the logit model: it allows us to test for PH violations more easily. As we described earlier, a testing procedure for logit exists, but is cumbersome. When we test for PH violations in both the Cox and the logit, we find a number of violations.[21] Table 3's third column estimates a Cox model with corrections for these PH violations. This model's results change the naïve estimates presented in the

[21] See Appendix D for the PH test results.

previous two columns. Economic interdependence's estimate is negative and large, indicating that greater levels of economic interdependence greatly reduce the risk of MID onset. However, economic interdependence's time interaction is positively signed. The constituent and interaction term together indicate that immediately following a MID ($t = 0$), higher levels of economic interdependence greatly reduce the risk of another MID onset. However, as time passes, this pacifying effect grows smaller and smaller. This is an important scope condition on a core finding in the study of MID onset: greater levels of economic interdependence do reduce the risk of MID onset, but this effect is strongest in the immediate aftermath of a MID and tends to decline over time.[22] The example underscores the substantive importance of testing for PH violations in any setting, regardless of estimator.

2. Causal Complexity

An additional advantage of Cox models compared to logit models is their flexibility in modeling political processes characterized by causal complexity (Metzger and Jones 2016). One particular form of causal complexity in BTSCS settings relates to properly defining the outcome of interest. As McGrath (2015) describes, many studies' datasets fail to differentiate between the *onset* of an event, and the subsequent *duration* of an event. To provide a concrete example, consider a MID that commences in 2000 and terminates in 2003. The incumbent practice among researchers is to code years in which MIDs are ongoing as 0. In this case, 2000 would be coded as 1, and 2001, 2002, and 2003 would be coded as 0. By doing so, researchers introduce a situation of heterogeneous 0's, where 0 may either indicate peace or an ongoing conflict. The complex effect of a particular factor, such as economic interdependence, is lost as its effect on conflict—both onset and duration—is simply assumed to be the same. The same principle holds if the dependent variable is coded as 1 in 2000 and each of the intervening years through 2003. The determinants of MID onset may differ from the determinants of MID duration, but by coding 2000-2003

[22] Others have worried about economic interdependence, temporal dependence, and how the effect of the former might be impacted by the latter. See, e.g., Oneal and Russett (1999).

in the same fashion, we introduce heterogeneity for the 1's, and different covariate effects for onset vs. duration are overlooked (McGrath 2015).

Cox models' flexibility affords researchers an elegant solution, one that directly models causal complexity while avoiding the pitfalls described by McGrath (2015). Specifically, an extension of Cox models, known as multi-state duration models (Metzger and Jones 2016), allow researchers to disaggregate complex political processes into discrete stages, and model the determinants of transitions between these stages. For MIDs, the process can be divided into two stages: peace and an ongoing MID. Transitions between these two stages are denoted by the occurrence of one of two events: MID initiation and MID termination, respectively. By disaggregating the MID process into these discrete stages, we can then account for causal complexity by using transition-specific covariates, which allow covariate effects to vary depending on the transition in question. Thus, if the same variable of interest exerts a one effect on the MID onset but another, different effect on MID termination, multi-state models will detect and model the difference.

Structuring the data to account for a situation with two stages and two possible transitions is straightforward. The dataset requires (1) a stage identifier, recording whether a dyad is currently involved in a MID or not, as well as (2) a transition indicator, denoting when transitions between stages occur. Here, it amounts to a variable coded 1 for dyad-years in which MID initiation or termination occurs. We can then generate transition-specific covariates, allowing the covariates' effect to vary across transitions.[23] With this data structure, any standard statistical package can estimate a multi-state model: estimate a Cox model, include the transition-specific covariates, and stratify the baseline hazards by stage.[24] Finally, both R and Stata have post-estimation utilities for calculating transition probabilities from the estimated multi-state model (for R: Wreede, Fiocco, and Putter 2010, 2011).

[23] For more on generating transition-specific covariates, as well as structuring data for more complex multi-state models, see Metzger and Jones (2016).

[24] Stratifying the baseline hazards according to whether a dyad is currently experiencing a MID or not accounts for the different underlying rates of MID onset and termination.

TABLE 4. Multi-State Model of MID Onset and Termination

	Peace → MID	MID → Peace
Allies	-0.290*†	0.185*
	(0.131)	(0.076)
Capability Ratio (ln)	-0.166**	0.070**
	(0.022)	(0.022)
Economic Interdependence (Low)	-23.23**	12.62**
	(6.000)	(3.42)
Contiguity	2.237**	-0.024
	(0.145)	(0.075)
Distance (ln)	-0.296**	-0.079*
	(0.047)	(0.038)
Major Power Dyad	1.589**	0.029
	(0.112)	(0.076)
Democracy (Low)	-0.046**	-0.004
	(0.010)	(0.006)
Joint IGOs	0.034**	0.002
	(0.004)	(0.002)
Allies * ln(Time)	0.135*	
	(0.056)	
Contiguity * ln(Time)	-0.162*	
	(0.064)	
Distance * ln(Time)	-0.112**	0.104**
	(0.024)	(0.035)
Major Power Dyad * ln(Time)	0.214**	
	(0.045)	
Democracy * ln(Time)	-0.009*	
	(0.004)	
Joint IGOs * ln(Time)	-0.012**	
	(0.002)	
Log-Likelihood (partial)	-18045.94	

† = $p \leq 0.10$, * = $p \leq 0.05$, ** = $p \leq 0.01$, two-tailed tests.

Table 4 presents the multi-state model results. The first column contains the estimates for each covariate's effect on MID onset, and the second column contains the estimates for MID termination, equivalent to a transition from an ongoing MID to peace. We tested for PH violations for both transitions, and implemented corrections accordingly. Table 4's results demonstrate the importance of accounting for causal complexity. Consider regime type's effect when operationalized as the least democratic state within the dyad. As this state becomes more democratic, the risk of MID onset declines (Peace → MID). With the time interaction, this effect grows increasingly negative, meaning that

democracy plays an even greater role in reducing the risk of MID onset as peace endures. However, regime type has no significant effect on MID duration (MID → Peace). Both results together indicate democracy may be helpful in avoiding MIDs in the first place, but if a MID does start, democracy plays little role in reducing the MID's duration. Without differentiating between these two distinct transitions, regime type's different effects might be overlooked. However, Cox models' flexibility allows us to straightforwardly evaluate this and many more causally complex hypotheses that would be significantly more difficult to test with logit.

V. Conclusion

We offer a new way to substantively interpret estimates from duration models, in the form of transition probabilities. Our proposed quantity has an easy interpretation, one arguably more straightforward than many current interpretation techniques. Further, they bridge the divide between modes of interpretation more broadly familiar to political scientists, such as predicted probabilities, and the less familiar modes of interpretation that accompany duration models. Transition probabilities view the process of interest as having at least two discrete stages (e.g., alive, dead). They subsequently tell us the probability that a subject occupies any particular stage by time t, given a starting time and starting stage. Current interpretation strategies in duration models fix these starting conditions by default—all subjects begin in the same stage and have just become at risk (implying $t = 0$). However, transition probabilities permit these conditions to be altered, providing researchers with more freedom to specify scenarios best aligning with their substantive hypotheses. In short, transition probabilities tell us about 'which' stage a subject likely occupies. To estimate transition probabilities, practitioners can use the `mstate` package in `R` (non-, semi-parametric models), our own package in Stata (non-, semi-parametric), or Crowther and Lambert's Stata package (non-parametric, parametric).

Transition probabilities have further implications for any political scientist working with BTSCS data. Typically, researchers opt for logit/probit because the models produce easily interpretable post-estimation quantities, compared to duration models. Transition probabilities nullify this line of

argumentation. We suggest this matters because, in some cases, duration models can more adroitly model certain processes than L/P can. We specifically highlighted the semi-parametric Cox model's ability to address three things: easily testing and correcting for PH violations, removing concerns about baseline hazard misspecification, and easily modeling causally complex processes. We used applications from judicial politics and interstate conflict to showcase these facets. In the process of doing so, we showed how transition probabilities provide useful, straightforward interpretations from Cox models.

Works Cited

Aalen, Odd, Ornulf Borgan, and Hakon Gjessing. 2008. *Survival and Event History Analysis: A Process Point of View*. New York: Springer.

Andersen, Per Kragh, Ornulf Borgan, Richard D. Gill, and Niels Keiding. 1993. *Statistical Models Based on Counting Processes*. New York: Springer.

Beck, Nathaniel. 2010. "Time is Not A Theoretical Variable." *Political Analysis* 18 (3): 293–294.

Beck, Nathaniel, Jonathan Katz, and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42 (4): 1260–1288.

Beyersmann, Jan, Arthur Allignol, and Martin Schumacher. 2011. *Competing Risks and Multistate Models with R*. New York: Springer.

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

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press.

Carter, David B., and Curtis S. Signorino. 2010. "Back to the Future: Modeling Time Dependence in Binary Data." *Political Analysis* 18 (3): 271–292.

Crowder, Martin J. 2012. *Multivariate Survival Analysis and Competing Risks*. Boca Raton, FL: Chapman and Hall/CRC.

Crowther, Michael J., and Paul C. Lambert. Forthcoming. "Parametric Multi-State Survival Models: Flexible Modelling Allowing Transition-Specific Distributions with Application to Estimating Clinically Useful Measures of Effect Differences." *Statistics in Medicine*.

Dabrowska, Dorota. 1995. "Estimation of Transition Probabilities and Bootstrap in a Semiparametric Markov Renewal Model." *Journal of Nonparametric Statistics* 5 (3): 237–259.

Esarey, Justin, and Jane Lawrence Sumner. 2016. "Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate." Working paper. http://jee3.web.rice.edu/research.htm.

Gill, Richard D., and Soren Johansen. 1990. "A Survey of Product-Integration with a View Toward Application in Survival Analysis." *The Annals of Statistics* 18 (4): 1501–1555.

Hall, Matthew E. K., and Joseph Daniel Ura. 2015. "Judicial Majoritarianism." *Journal of Politics* 77 (3): 818–832.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–161.

Hougaard, Philip. 2000. *Analysis of Multivariate Survival Data*. New York: Springer.

Jones, Bradford S., and Regina P. Branton. 2005. "Beyond Logit and Probit: Cox Duration Models of Single, Repeating, and Competing Events for State Policy Adoption." *State Politics & Policy Quarterly* 5 (4): 420–443.

Kalbfleisch, John D., and Ross L. Prentice. 2002. *The Statistical Analysis of Failure Time Data*. 2nd ed. Hoboken, NJ: Wiley-Interscience.

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): 347–361.

Klein, John P., and Melvin L. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer.

Kropko, Jonathan, and Jeffrey J. Harden. 2015. "Beyond the Hazard Ratio: Generating Expected Durations from the Cox Proportional Hazards Model." Presented at the International Methods Colloquium, 16OCT.

Licht, Amanda A. 2011. "Change Comes with Time: Substantive Interpretation of Nonproportional Hazards in Event History Analysis." *Political Analysis* 19 (2): 227–243.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

McGrath, Liam F. 2015. "Estimating Onsets of Binary Events in Panel Data." *Political Analysis* 23 (4): 534–549.

Metzger, Shawna K., and Benjamin T. Jones. 2016. "Surviving Phases: Introducing Multistate Survival Models." *Political Analysis* 24 (4): 457–477.

Mills, Melinda. 2011. *Introducing Survival and Event History Analysis*. Los Angeles: Sage.

Oneal, John R., and Bruce Russett. 1999. "The Kantian Peace: The Pacific Benefits of Democracy, Interdependence, and International Organizations, 1885-1992." *World Politics* 52 (1): 1–37.

------. 2005. "Rule of Three, Let It Be? When More Really Is Better." *Conflict Management and Peace Science* 22 (4): 293–310.

Putter, H., M. Fiocco, and R. B. Geskus. 2007. "Tutorial in Biostatistics: Competing Risks and Multi-State Models." *Statistics in Medicine* 26 (11): 2389–2430.

Singer, Judith D., and John B. Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford: Oxford University Press.

Wreede, Liesbeth C. de, Marta Fiocco, and Hein Putter. 2010. "The mstate Package for Estimation and Prediction in Non- and Semi-Parametric Multi-State and Competing Risks Models." *Computer Methods and Programs in Biomedicine* 99 (3): 261–274.

------. 2011. "mstate: An R Package for the Analysis of Competing Risks and Multi-State Models." *Journal of Statistical Software* 38 (7): 1–30.

Appendix A:
Cox/cloglog Equivalence Derivation

In this appendix, we reproduce the derivation demonstrating that a continuous-time Cox semi-parametric duration model is equivalent to a discrete-time cloglog model with time dummies for its temporal dependence correction (also referred to as "grouped duration data"). We include the equivalence derivation for reference purposes; others have shown the same elsewhere (e.g., Beck, Katz, and Tucker 1998, 1284–1285; Cameron and Trivedi 2005, 600–601, 602–603). We focus on the general intuition behind the derivation, leaving the underlying mathematics to the previously cited others. We try to use *j* to denote calendar time (e.g., year) for an observation, and continue using *t* for the duration.

The place to begin is by realizing that a continuous Cox model is usually expressed in terms of $h(t)$, the hazard, and that a cloglog model (discrete time or otherwise) is usually expressed in terms of $\Pr(y = 1)$, the probability of an event occurring. When we speak of discrete-time durations, $y_{ij} = 1$ means the duration terminates in time period *j* for subject *i*. The cloglog's functional form is (Long 1997, 51):

$$\Pr(y = 1) = 1 - \exp[-\exp(\beta' X)] \qquad 6$$

To show equivalence, we will need to first reexpress the continuous-time Cox's hazard function in terms of a probability. We can then convert the continuous-time Cox expression into discrete time, and see if the resultant expression is equivalent to equation 6.

1. Continuous Cox: Hazard to Probability

In the main text, we already discussed how to express the Cox hazard in terms of a probability. We start from $h(t)$:

$$h(t) = h_0(t) \exp\left(\beta' X_{ij}\right) \qquad 7$$

From $h(t)$, we can derive the expression for $S(t)$, the survivor function, because writing an expression for $h(t)$ necessarily specifies a distribution for *t* (Kalbfleisch and Prentice 2002, 7). If we know *t*'s distribution, any function involving *t*'s distribution—its probability density (PDF), cumulative

distribution (CDF), survivor, hazard, or cumulative/integrated hazard—can be expressed in terms of the other functions. $h(t)$ and $S(t)$'s connection is through the basic identity:

$$S(t) = \exp\left(-\int_0^t h(u)du\right) \quad 8$$

where u denotes all times at which we observe any failure event on the interval $(0,t]$.[25] $S(t)$ tells us the probability that a subject has *not* failed by time t. For our probability, we want to know whether the subject *has* failed by time t. This quantity is equal to the complement of $S(t)$, $1 - S(t) \equiv F(t)$, t's cumulative distribution function.

$$\Pr(y = 1) = F(t) = 1 - \exp\left(-\int_0^t h(u)du\right) \quad 9$$

We now have our continuous-time Cox model expressed in terms of a subject's probability of failing by t.

2. Continuous Time to Discrete Time

Table 5. Data Structure for Two Subjects

Continuous Time						Discrete Time					
i	j_0	j	t_0	t	y	i	j_0	j	t_0	t	y
1	1990	1993	0	4	1	1	1989	1990	0	1	0
2	2005	2007	0	3	1	1	1990	1991	1	2	0
						1	1991	1992	2	3	0
						1	1992	1993	3	4	1
						2	2004	2005	0	1	0
						2	2005	2006	1	2	0
						2	2006	2007	2	3	1

Note: j_0 = first year where i is at risk. y = 1 if subject fails by t, 0 otherwise.

We use "discrete time" to refer to situations in which a subject experiences failure sometime in a fixed interval, but we do not observe precisely when.[26] For instance, if we are recording BTSCS data on an annual basis, y_{ij} will tell us that our failure event happened sometime between the end of year $j - 1$ and the end of year j—$(j - 1, j]$. Using a discrete-time model for BTSCS data tells us the probability of our

[25] $\int_0^t h(u)du = H(t)$, the cumulative hazard function. The cumulative hazard's value in t will be the sum of all hazard values for every time point from 0 up through t.

[26] This particular situation is sometimes referred to "grouped duration data" instead of discrete-time data.

event occurring sometime during year j. We compute t, the duration, by counting the number of years in which the subject has been at risk but not failed (or the number of years since the subject's last failure, if subjects can experience the failure event multiple times). Once at risk, the first year without a failure would be $t = 1$, the second year, $t = 2$, and so on.

We acknowledge the shift to discrete time by modifying the interval over which we integrate the hazard. To explain by way of example: Table 5 shows data on two subjects, in a continuous-time format and a discrete-time format. Both formats use counting-process notation by adding a column for t_0 (Box-Steffensmeier and Jones 2004, 99–101). t_0 represents where the previous observation "left off" counting for t within a panel, and t continues to represent "up through the end of this time period," also within the panel. t's counting-process interpretation is consistent with our previous discussion of t in discrete time—it represents the probability of our event occurring sometime during year j (t), but after year $j - 1$ (t_0). For continuous time, the interval's starting point was implicitly $t_0 = 0$,[27] giving us our integration interval of $(t_0,t] = (0,t]$. For discrete-time data (e.g., annual), $t_0 = t - 1$. Our integration interval shifts for discrete-time data to reflect each observation's t_0. Equation 9, for discrete time, becomes:

$$\Pr(y_{ij} = 1) = 1 - \exp\left(-\int_{t-1}^{t} h(u)du\right) \qquad 10$$

If we insert equation 7 into equation 10 and simplify, we get:

$$\begin{aligned} \Pr(y_{ij} = 1) &= 1 - \exp\left(-\int_{t-1}^{t} \left[h_0(u)\exp(\beta' X_{ij})\right]du\right) \\ \Pr(y_{ij} = 1) &= 1 - \exp\left(-\exp(\beta' X_{ij})\int_{t-1}^{t} [h_0(u)]du\right) \end{aligned} \qquad 11$$

The final piece to recognize is that $\int_{t-1}^{t}[h_0(u)]du$ will return different values <u>across</u> t's, but will be constant <u>within</u> a $(t - 1, t]$ interval. Let $\alpha_t = \int_{t-1}^{t}[h_0(u)]du$, to reinforce that the integral's value is constant within each time interval. If we substitute α_t into equation 11 and rearrange terms, we obtain:

[27] $t_0 = 0$ implies no left truncation for either subject.

$$\Pr(y_{ij} = 1) = 1 - \exp\left(-\exp(\beta' X_{ij})\,\alpha_t\right)$$

$$\Pr(y_{ij} = 1) = 1 - \exp\left(-\exp\left(\beta' X_{ij} + \ln(\alpha_t)\right)\right) \quad (12)$$

Moving α_t inside the parenthesis containing $\beta' X_{ij}$ makes clear that we are simply adding some constant—because the natural log of a constant is another constant—to the regression line (represented by $\beta' X_{ij}$). Adding a constant to the regression line changes the line's intercept by shifting it up or down. Since $\ln(\alpha_t)$ is indexed by *t*, the intercept shifts effectively give each *t* its own fixed effect. We can express fixed effects for *t* by including time dummies (denoted $\tau_t = \ln(\alpha_t)$ by BKT), giving us a temporal dependence correction. Equation 12 is identical to equation 6 if we added time dummies to the latter.

Appendix B:
Brief Duration Model Overview

Duration models are concerned with lengths of time as dependent variable. They are useful when we are investigating questions about how long before a subject experiences some event. We count the number of periods that the subject "survives" before experiencing the event; the resultant quantity is t, our duration of interest and dependent variable. OLS is not particularly well-suited to model durations, because it assumes (1) the dependent variable can be positive or negative, (2) that the regression equation's errors are normally distributed, and (3) that all our subjects fail during our observation period, meaning that we know the start and end point of every subject's duration (Crowder 2012, 3–5). One or more of these assumptions are false when working with durations as a dependent variable—quantities of time are non-negative, the errors they subsequently produce can be non-normal, and we do not always observe the duration's end point for all our subjects (i.e., we may have right-censored durations).

Duration models come in three major classes. The classes are differentiated from one another based on two factors. The first factor is *baseline hazard functional form*. It refers to whether we assume a functional form connecting the baseline hazard (h_0) to our observed t's, where the baseline hazard expresses the hazard of the duration's terminating event occurring when there are either no covariates in the model, or when all of the covariates are set to zero. The second factor is the covariate link function. Here, we refer to whether there are covariates in the model, because if there are, we must specify a function connecting these covariates to our observed t's. We use x to refer to a single covariate, and X to refer to the entire set of included covariates.

We begin our discussion with continuous-time durations, meaning that our dataset has one record for every subject we observe. If we assume no functional form for the hazard and have no covariates, we obtain a non-parametric duration model. The Kaplan-Meier curve is the classic example, and is usually expressed in terms of the survivor function $S(t)$, which represents the fraction of subjects still at risk in t.[28] Its equation is (Klein and Moeschberger 2003, 92):

[28] If a semi-parametric duration model does not include covariates, the resulting $h(t)$ is a non-parametric expression.

$$S(t) = \prod^{t < t_{\text{Max}}} \left(1 - \frac{d_t}{N_t}\right) \quad 13$$

implying that the hazard function is:

$$h(t) = \frac{d_t}{N_t} \quad 14$$

where t_{Max} represents the largest observed failure time, d_t represents the number of subjects that fail in time period t, and N_t represents the number of subjects still at risk at the start of t, which some denote as N_{t-}.

If we continue to impose no functional form for the baseline hazard, but include covariates in our specification, we now have a semi-parametric duration model. Cox proportional hazard models are the most ubiquitous of the semi-parametric duration models. Cox models take the form:

$$h(t) = h_0(t)\exp(\beta' X) \quad 15$$

Notice how the covariates (X) and their corresponding estimates (β) are linked to $h(t)$ using an exponential link function—we are exponentiating the linear score of β and X's product.

Finally, if we do make a functional form assumption about the baseline hazard, with or without covariates, we obtain a parametric duration model. There are many variants of parametric duration models. They differ from one another in terms of the baseline hazard's functional form. For example, a Weibull parametric duration model, when specified in terms of proportional hazards (denoted with "PH" subscript), has the form:

$$h(t) = pt^{p-1}\exp(\beta_{PH}{}' X) \quad 16$$

where p is the hazard's shape parameter, a quantity to be estimated from the data. Box-Steffensmeier and Jones (2004, chap. 3) provide a list of parametric duration models commonly used in political science, along with the models' functional forms.

On a concluding note, duration models' estimated coefficients can come in one of two metrics: accelerated failure time (AFT) or proportional hazards (PH). The coefficients' metric matters because a positive AFT coefficient means the *opposite* of a positive PH coefficient. Some duration models'

coefficients can be reported in either metric (e.g., Weibull, exponential),[29] but some models' coefficients can only be reported in AFT (e.g., generalized Gamma, log-logistic, log-normal) or only in PH (e.g., Gompertz, Cox). Coefficients in an AFT metric speak directly to *x*'s effect on the *duration*, where a positive β_{AFT} means higher levels of *x* will lengthen *t*, the duration of interest. By contrast, coefficients from PH metric speak in to *x*'s effect on the *hazard* of the duration ending ($h(t)$). Specifically, the hazard acts like a conditional probability, modeling the probability that subject *i* will experience the event terminating our duration in time period *j*, given that *i* has not experienced the event yet. A positive β_{PH} means higher levels of *x* increase the probability of the duration's terminating event, implying the duration itself will be shorter.

[29] Importantly, the AFT vs. PH metrics for Weibull and exponential parametric models are based on the same underlying math. You can transform AFT coefficients into PH coefficients or vice-versa (Box-Steffensmeier and Jones 2004, 29).

Appendix C:
Simulation Procedure – Outline

1. Decide on a number of subjects to move through the process, a starting time (s), a starting stage, and an end time (t) for all subjects.
2. For each subject:
    a. Set s to the current time and the starting stage as the current stage.
    b. Ensure the subject is at risk of a transition. If it is not (e.g., it is in an absorbing stage), move to the next subject.
    c. From the set of all observed failure times, randomly select a transition time larger than the current time. Call this t^*.
    d. If $t^* > t$, subject stays in current stage until end of specified time interval t. Move to the next subject.
    e. Otherwise, select which transition the subject will experience, with each transition's selection probability being equal to $h(t^*)$.
    f. Record t^* as the new current time and (2e)'s stage as the new current stage, for clock time durations. Gap time is the same, except that the new current time resets to 0.
    g. Repeat (2b)-(2f) until the subject's $t^* > t$ or until the subject enters a stage with no outward transitions.
3. From the simulation results, count the number of subjects occupying each stage at each unit interval in $(s,t]$.
4. Loop over 1-3 for the desired number of simulation runs.

Appendix D:
Proportional Hazards Tests for Logit Models

The procedure for testing for violations of the proportional hazards assumption in the context of panel logit models is described by Carter and Signorino (2010). Testing for violations of the PH assumption consist of performing a series of likelihood-ratio tests. The restricted model in the tests is the base specification of the model, which assumes proportional hazards. The unrestricted specification(s) are similar to the base model, but include an interaction between a covariate and the three time polynomials. For each covariate in the model, it is necessary to estimate a unique model with time interactions. Once these models have been estimated, it is then possible to perform a series of likelihood-ratio tests of equivalence between the core restricted model and each of the unrestricted models. The result of these models indicates whether the interactions between each covariate and the time polynomials significantly improves model fit. If the test result is statistically significant, it indicates a likely violation of the PH assumption for that particular covariate. To correct for these violations, as with the Cox model, it is possible to include an interaction between the violating covariate and each of the terms in the time polynomial.

Table 6 presents the results of tests for violations of the PH assumption for the logit model in Table 3 of the main text. Each row contains the result of a likelihood-ratio test of the core restricted model, and an unrestricted specification of the model including an interaction between the covariate and each of the three time variables. Table 6 indicates that each of the covariates in the model likely violates the PH assumption, as the test statistic for each likelihood-ratio test is significant at the $p < 0.000$ level. To correct for these violations, it is necessary to include interactions between each of the covariates in the model, and each of the time terms. This result is the same as that attained from tests of Schoenfeld residuals following estimation with a Cox model.

TABLE 6. Proportional Hazards Tests for Logit Specification

	χ^2	*p*-value
Allies	26.06	0.000
Capability Ratio (ln)	45.57	0.000
Economic Interdependence (Low)	47.75	0.000
Contiguity	93.90	0.000
Distance (ln)	19.17	0.000
Major Power Dyad	18.76	0.000
Democracy (Low)	46.97	0.000
Joint IGOs	89.38	0.000

Test statistics result from a likelihood-ratio test of equivalence between the base restricted model, and an unrestricted variant with time interactions for each covariate.