# Different Words, Same Song: Advice for Substantively Interpreting Duration Models

**Benjamin T. Jones,** *University of Mississippi*
**Shawna K. Metzger,** *College of William & Mary*

**ABSTRACT** The use of duration models in political science continues to grow, more than a decade after Box-Steffensmeier and Jones (2004). However, several common misconceptions about the models still persist. To improve scholars' use and interpretation of duration models, we point out that they are a type of regression model and therefore follow the same rules as other more commonly used regression models. In this article, we present four maxims as guidelines. We survey the various duration model interpretation strategies and group them into four categories, which is an important organizational exercise that does not appear elsewhere. We then discuss the strengths and weaknesses of these strategies, noting that all are correct from a technical perspective. However, some strategies make more sense than others for nontechnical reasons, which ultimately informs best practices.

Political scientists are no strangers to duration models, which allow researchers to test hypotheses about how long until an event of interest occurs. Box-Steffensmeier and Jones' (2004) work on duration models arguably marks the watershed moment for these models' use in political science. The main virtue of the models is the ease with which they handle potential duration dependence—formally, that the likelihood of an event may be contingent on how long a subject has been at risk.[1]

We both work frequently with duration models; accordingly, we happily field duration model–related queries from our colleagues and students. In doing so, we were struck by how often we receive the same questions, which cluster into the following three groups:

1. People asking how duration models "work" and then being surprised when we use other models to explain (e.g., logit or probit).
2. People asking how to interpret duration models. Usually, this manifests as (1) a preoccupation with one specific interpretation strategy, without seeing how different interpretation methods are related; or (2) being overwhelmed with the number of possible interpretation strategies and lacking a clear sense of where to begin.
3. People asking how to compute a duration model quantity in R or Stata.

With these frequent queries in mind, we searched for a piece or two that succinctly organized and summarized our answers in one place. To our surprise, no such piece existed.

This article synthesizes our more frequent answers. Our answers' overarching theme is that duration models are a type of regression model and, as such, most of the general intuition and best practices gleaned from linear regression models, logit models, and others apply equally. Duration models have an extra "wrinkle" or two because they can address right-censored data; however, these wrinkles are accommodated automatically when using the models. There is nothing otherwise unique or special about duration models' underlying principles that should lead practitioners to jettison their intuition when estimating them.[2] Appreciating this point is important because many of the best practices and rules of thumb internalized by practitioners in the context of other regression models are inconsistently heeded in the context of duration models. As a notable example, reporting *p*-values or confidence intervals around predicted probabilities from a logit or probit model is ubiquitous, whereas reporting similar measures of uncertainty from a duration model is inconsistent at best, as we illustrate with two meta-analyses.

**Benjamin T. Jones** *is assistant professor of political science at the University of Mississippi. He can be reached at btjones1@olemiss.edu.*
**Shawna K. Metzger** *is visiting assistant professor of government at the College of William & Mary. She can be reached at shawna@shawnakmetzger.com.*

doi:10.1017/S104909651900060X  

Not recognizing the connection between duration and other regression models also has negatively affected whether and how scholars *interpret* their duration model results. For instance, some researchers simply stop after interpreting the model's estimated coefficients. When substantive interpretation does happen, practitioners use various strategies with varying degrees of success. The plethora of interpretation strategies seems to deepen the uncertainty and mystery surrounding the models.

To engage with these issues, we present four stylized maxims about interpreting duration models that represent major areas where practitioners' use of them might go awry. We articulate these maxims by drawing parallels to more widely used regression models to emphasize our central point: the hard-won intuition that practitioners have developed with other regression models applies equally to duration models. Our maxims advise practitioners to move beyond the regression table when interpreting duration models. We categorize existing duration model interpretation strategies to provide practitioners with a concise overview—an important organizational exercise that does not exist elsewhere.[3] Following this overview, we encourage practitioners to use their paper's substance and theory when deciding which interpretation strategy to employ while also underscoring the universal importance of measures of uncertainty. We conclude by assessing the various interpretation strategies' strengths and weaknesses in terms of presenting and interpreting the results, followed by providing more general guidance for how to use these interpretation strategies in conjunction with one another to maximize their effect.

*Our maxims advise practitioners to move beyond the regression table when interpreting duration models. We categorize existing duration model interpretation strategies to provide practitioners with a concise overview—an important organizational exercise that does not exist elsewhere.*

## THE MAXIMS

### #1: You Cannot Directly Interpret the Coefficients as Substantive Effects

Regardless of which duration model you estimate,[4] all duration models with covariates are nonlinear in parameters—the same as logits, probits, and count models, among others. Therefore, we cannot directly interpret the magnitude of any $\beta$s as we might in a simple additive linear regression model because they are not equivalent to the coefficients' substantive effects (marginal or otherwise) on $y$. Instead, we must generate additional quantities to present our model's substantive results (King, Tomz, and Wittenberg 2000).

Stated simply, you cannot stop at the regression table. Our hypotheses are usually about $x$'s effect on $y$. However, in all nonlinear models, $\beta$ does not tell us about the relationship between $x$ and $y$ but rather between $x$ and $y^*$.[5] To reach a conclusion about $y$, we need to convert $y^*$ back to the quantity we care about, $y$, through a link function. In a logit, applying the logistic link transforms $y^*$ into a new quantity, $y$ (technically, $\Pr(y{=}1)$), whose values fall between 0 and 1, yielding probabilities.[6] For duration models, the usual link function is $\exp(y^*)$, which produces a $y$ ($\equiv t$) that must be greater than 0, as time cannot be negative. Without this link function, $\beta$ represents $x$'s effect on either the log-hazard (for proportional hazard models) or the log-duration (for accelerated failure time models),[7] neither of which is likely to be the focus of our hypotheses. Therefore, you must transform the $\beta$s in some way to glean substantive meaning, as discussed in Maxim #2. *How* you should transform the $\beta$s is directly related to your hypothesis about $x$'s effect on $y$—specifically, the way in which you framed your discussion of $y$, as we discuss further in Maxim #3.

### #2: To Generate Substantive Effects, You Will Need to Do "Something" to the Coefficients

There are several ways to substantively interpret duration-model results. Each interpretation is technically correct (see Maxim #3), which is both a blessing and a curse—no matter which strategy you choose, your inferences will be *technically* sound, but clear-cut standards *cannot* exist regarding when some strategies perform better than others. Therefore, determining the best strategy for your work entails getting a sense of what the different strategies are, how they relate to one another, and which questions a respective strategy allows you to address most easily.

We loosely group extant interpretation strategies along two dimensions. The first relates to the underlying quantity of interest. The second dimension pertains to whether the quantity is an absolute or a relative quantity. The result is four groupings, shown in table 1.

There are four noteworthy observations from the table:

1. All interpretation strategies involve either exponentiating the model's $\beta$s or generating a predicted quantity using the $\beta$s, for reasons discussed in Maxim #1.
2. Duration models have two families of predicted quantities. These focus on how long until something occurs (i.e., the duration, $t$) or, equivalently, on the risk that something occurs (i.e., the hazard,

*Table 1*

**Current Approaches to Substantive Interpretation**

| | | Absolute (Levels) | Relative (Difference) |
|---|---|---|---|
| Quantity | $t$ | Mean or median duration [$E(t)$ or $Q_{50}(t)$] | Time ratio, marginal effect, first difference [$\exp(\beta_{AFT})$, $\partial t/\partial x$, $\Delta E(t)$, or $\Delta Q_{50}(t)$] |
| | Risk | Hazard rate, cumulative hazard, survivor, transition probabilities [$h(t)$, $H(t)$, $S(t)$, $\Pr(g,h)$ in $(s,t]$] | Hazard ratio, % change in hazard rate [$\exp(\beta_{PH})$, $\%\Delta h(t)$] |


https://doi.org/10.1017/S104909651900060X Published online by Cambridge University Press

$h(t)$); $t$ and $h(t)$ are the duration model equivalent of OLS's predicted $y$ ($\hat{y}$).[8] For exposition purposes, think of hazards as being conceptually similar to probabilities, but also note the two are *not* usually synonyms.

3. Parametric duration models are built from asymmetric distributions, which means the mean and median survival times will not be equal. Observed survival times tend to be right-skewed, and the implications of computing any skewed variable's mean versus median apply equally in a duration context. Furthermore, there may be right-censored subjects—subjects that will eventually experience the event of interest but have not experienced it *yet* when last observed.[9] Typically, right-censored observations fall in the right tail of $t$'s observed distribution.[10] As a general rule of thumb, the more right-censored subjects there are, the more appealing the median becomes.
4. There are numerous ways to quantify "risk" from a duration model, including the risk of an event occurring (the hazard, $h(t)$); the probability of an event *not* occurring (the survivor function, $S(t)$); the total risk that an event will have occurred (the cumulative hazard function, $H(t)$); and the probability that an event will have occurred (transition probabilities). Each of these is a different way to express the same underlying concern: how (un-)likely is it that an event will occur by some point in time.

Some of these quantities are easier to compute than others, depending on users' statistical program of choice. We inventory R and Stata's respective capabilities in appendix D.

### #3: All of the Techniques Are Correct from a Technical Perspective, but Some Make More Sense Than Others

All of these interpretation strategies come from the same underlying model estimates (i.e., the $\beta$s and standard errors). Thus, *all* of these strategies are correct, from a technical perspective,[11] the same way that odds ratios and predicted probabilities are equally correct ways of interpreting logit output. Therefore, you must use other nontechnical criteria to guide decisions about which strategy to employ. We suggest considering two criteria, both relating to your paper's presentation.

First, consider the extent to which a given interpretation strategy matches your theory and hypotheses. Kropko and Harden (forthcoming) make this point succinctly: If your hypothesis is framed in terms of durations—for instance, how long until a civil war recurs—then presenting your results in terms of durations logically follows (e.g., mean or median duration). Conversely, if your hypothesis is framed in terms of events (e.g., which factors make a civil war more likely to recur), then presenting the results in terms of risk-based quantities would be a better match. These quantities speak more directly to an event's occurrence (or lack thereof) by depicting the conditions under which subjects are more likely to "survive"—here, that states remain at peace. Although our first point appears fairly simple and logical, misalignment between framing and interpretation strategies is rampant in political science. Kropko and Harden's (forthcoming) meta-analysis of 80 articles using Cox duration models reveals that approximately 33 (41.25%) have predominantly duration frames and another 10 to 14 use both frames equally (12.5%–17.5%). However, *none* of these articles generate duration-based quantities for interpretation.

Second, consider the relative ease with which a particular post-estimation quantity allows you to present and interpret your results. Some techniques may be more straightforward for your audience to understand than others. We return to this point in our broader discussion of the various quantities' strengths and weaknesses.

### #4: Whatever Technique You Choose, You Will Need *p*-Values or Confidence Intervals

With duration models, as with other regression models, measures of uncertainty around our predicted quantities improve our ability to make inferences. Standard practice for logit/probit models is to report such measures around both the estimated coefficients *and* any predicted quantities. However, standard practice for duration models is much less consistent. A review of all articles in the *Journal of Politics* in 2017 underscores this point. Of these articles, 10 report at least one logit or probit model in the main text, and 9 of those 10 analyses generate a post-estimation quantity[12] with some measure of uncertainty, suggesting that this practice is well internalized.[13] In comparison, four articles estimate duration models, with measures of uncertainty reported less consistently. They are reported for any first differences and hazard ratios but not for survival curves.

*This pattern holds more broadly: practitioners inconsistently report measures of uncertainty for duration model post-estimation quantities.*

This pattern holds more broadly: practitioners inconsistently report measures of uncertainty for duration model post-estimation quantities. We examine all articles in the *Journal of Politics*, *American Journal of Political Science*, and *American Political Science Review* from 2012 to 2016 and assess whether they report (1) a Cox model in the main text, and (2) any post-estimation quantities in the main text.[14] There are 16 such articles,[15] of which four do not report any post-estimation quantities (25%).

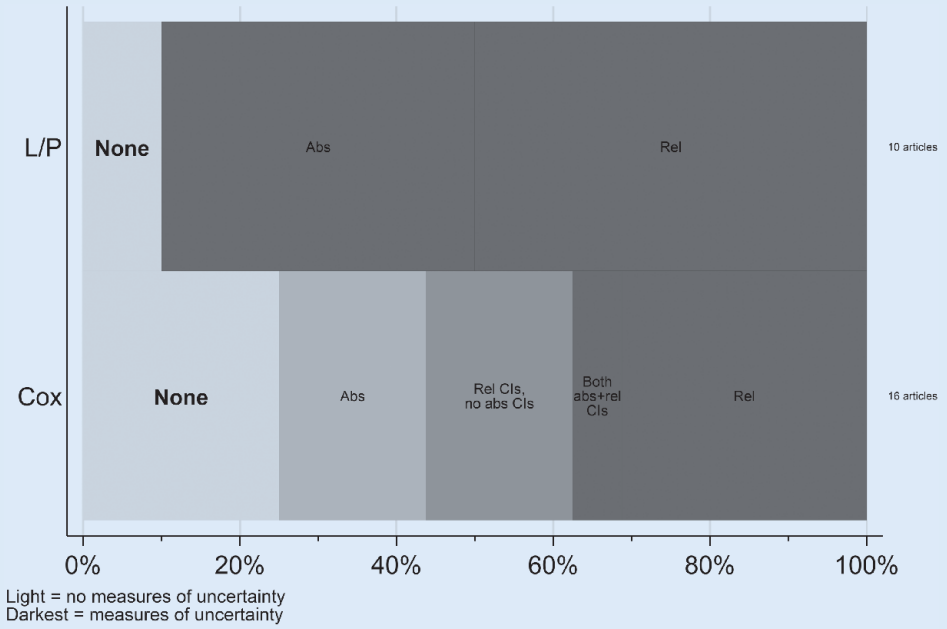
Of the remaining 12 articles that do report post-estimation quantities,[16] hazard ratios appear most frequently (9 of 12), each time with a measure of uncertainty. However, hazard ratios have weaknesses stemming from being a measure of relative change, as we elaborate on in the next section. Five of these 12 articles report only hazard ratios (Cox's "Rel" segment in figure 1), equivalent to a logit analysis reporting and interpreting odds ratios only. Finally, 6 of the 12 articles report a survivor curve ($S(t)$) and/or hazard rates ($h(t)$), but only one includes a measure of uncertainty around the quantity.[17]

Figure 1 visually depicts the two patterns from our two meta-analyses. The bars for both models should be filled entirely with the darkest gray if all articles report post-estimation quantities with measures of uncertainty. This is clearly not the case for duration models, indicating that the interpretation of duration

https://doi.org/10.1017/S104909651900060X Published online by Cambridge University Press

*Figure 1*

**Meta-Analyses Comparison**




Light = no measures of uncertainty
Darkest = measures of uncertainty

None: no post-estimation quantities reported
Abs: absolute post-estimation quantities (pred prob, $S(t)$, $h(t)$)
Rel: relative post-estimation quantities (odds ratios, marginal effects, first differences, % change in pred prob, hazard ratios)

Note: "CIs" used as shorthand for "any measure of uncertainty."

- Relative measures expressed in terms of ratios or percentages can be misleading. For instance, say that increasing $x$'s value produces a 100% increase in the hazard rate. However, a 100% increase could result if $h(t)$ increased in value from 0.4 to 0.8 (a fairly frequent event), but it also could result if the hazard increased from 0.00001 to 0.00002 (a very infrequent event). As Hanmer and Kalkan (2013, 265) point out, knowing something about the absolute level of the hazard, probability, or duration is "a necessary element for determining substantive significance." Yet, figure 1 illustrates that linking absolute and relative quantities in duration models is rare. More Cox model articles report only a relative measure compared to those that report both relative and absolute measures (total "Rel" segment size > total "Rel + Abs" segment size).

models differs from similar models. This trend is troubling because without measures of uncertainty, drawing meaningful inferences from post-estimation quantities can be difficult.[18]

## GENERAL STRENGTHS AND WEAKNESSES

Although all of the quantities in table 1 are correct, they are not equally useful in all cases. Each has strengths and weaknesses in terms of ease of interpretation, for both the researcher and the audience. We discuss individual strengths and weaknesses in appendix B, but summarize several broader rules of thumb here. As we noted earlier, most of the following rules of thumb apply to post-estimation quantities from any regression model, not only duration models, but researchers often overlook this similarity.

- If you generate absolute quantities, you will have to generate at least two covariate profiles[19] with different $x$ values. Otherwise, you will be unable to demonstrate how changes in $x$'s value bring about change in the predicted quantity—a necessity for substantive significance.
- Rainey (2017) points out that predicted quantities do not automatically inherit the $\beta$s' unbiased properties for any nonlinear regression model. Practitioners should be particularly mindful of biased predicted quantities when sample sizes are small.
- If you are looking at absolute quantities, the hazard and cumulative hazard are not scaled in especially intuitive units.[20] Comparatively speaking, duration-based quantities, the survivor, and transition probabilities have a far more intuitive scale, with the first scaled in the same units as the duration variable (e.g., months or years) and the survivor and transition probabilities expressed as probabilities.

Our advice is the same as Hanmer and Kalkan's (2013) for reported quantities. We prefer using *both* absolute and relative quantities in general, and duration models are no exception. We typically begin by mentioning whether the coefficient is statistically different from zero. We then move to absolute quantities to give readers information about the quantity's magnitude, calculating these absolute quantities for various covariate profiles of interest. We also usually check to see whether the profiles' confidence intervals overlap with one another—although with caution because overlapping confidence intervals do *not* necessarily mean a lack of statistical significance (Austin and Hux 2002; Bolsen and Thornton 2014; Schenker and Gentleman 2001).[21] Following this, we mention relative quantities to clearly and concretely communicate to readers the relative change in the quantity's value. By the end, readers have the information they require to make judgments about our results' substantive and statistical significance with relative ease.

## CONCLUSION

Duration models' usage has grown in political science, but researchers' adeptness with them has grown at a slower rate. Many applications have been limited by a lack of clear best practices for substantively interpreting the models' results. This article bridges these gaps by providing some rules of thumb to guide substantive interpretations of duration models. At its core, this set of maxims is built from a straightforward yet often underappreciated claim: duration models are like any other type of regression model with which political scientists work. As a result, almost all of the same guidance that political scientists receive with respect to interpreting a logit model, for instance, applies equally to duration models. Yet, our meta-analyses of


https://doi.org/10.1017/S104909651900060X Published online by Cambridge University Press

published articles illustrate that this guidance is not applied to duration models as frequently as other models. Overall, then, the message is clear: political scientists can do better when it comes to interpreting duration models.

## SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S104909651900060X

## ACKNOWLEDGMENTS

The authors' names appear in alphabetical order. Parts of this article were presented in the 2016–2017 International Methods Colloquium series and the University of Mississippi's International Relations Workshop. We thank Daniel Kent for feedback. We bear sole responsibility for any remaining errors and shortcomings. All analyses were performed using Stata 14.2 unless otherwise noted. ■

## NOTES

1. For more about the virtues of duration models, see Box-Steffensmeier and Jones (1997) and Metzger and Jones (2019).
2. Beck, Katz, and Tucker's (1998) canonical article also points out the strong linkages between duration models and logit/probit models: binary time-series cross-section data *are* grouped duration data. We also address and exploit this linkage in other work (Metzger and Jones 2019).
3. We provide a set of example interpretations with a substantive application in appendix C and a thorough list of the various R and Stata commands in appendix D.
4. Numerous options are available across three broad classes: parametric, semi-parametric, and non-parametric.
5. This is $y^*=XB$, which is referred to, equivalently, as the linear predictor or linear combination.
6. Berry, DeMeritt, and Esarey (2010) discuss this section's overall point in a logit/probit context.
7. See endnote 8 for an explanation.
8. The $t$ versus $h(t)$ distinction maps to the metric in which the specific duration model is expressed. Some duration models express their covariate effects in terms of the *duration* itself (i.e., accelerated failure time), whereas others are expressed in terms of the event's hazard—specifically, the event terminating the duration (i.e., proportional hazards). These different metrics constitute equally valid ways to make inferences.
9. Singer and Willett (2003, sec. 9.3) discuss "how and why censoring arise[s]" in a useful way with examples.
10. Importantly, this need not be so. Consider longitudinal medical studies, in which ill subjects receive some treatment and researchers then watch to see how long they stay healthy. We can lose subjects from wave to wave, for reasons unrelated to our process of interest. These subjects become right censored but have recorded durations less than the maximum.
11. Computing time-related predicted quantities from accelerated failure time (AFT) models in the presence of time-varying covariates—duration model speak for panel data in which covariates vary within panels, more often than not—can be laborious, mainly for conceptual reasons (Cleves et al. 2010, 241–44). There is no issue whatsoever with *estimating* a model in AFT with time-varying covariates.
12. Typically, these are predicted probabilities.
13. The same pattern holds if we expand to "any nonlinear-in-parameters model other than duration." We gain four additional papers—two ordered logits, one count model, and one hierarchical logit—of which three report a post-estimation quantity with confidence intervals. The fourth uses ordered logit and reported nothing additional.
14. We use Cox models for this meta-analysis because the search words are simpler than parametric duration models, in which several different phrases and combinations exist. Thus, we have more confidence that our set of Cox-related articles is complete, given the journals and years we used.
15. Park and Hendry (2015) also appears in this time frame. However, their article is entirely methodological and therefore does not contain post-estimation quantities. We removed this article from our sample, producing our final count of 16 articles.
16. The subsequent counts we mention sum to more than 12 because five articles reported multiple post-estimation quantities (segments labeled "rel + abs" in figure 1).
17. One of these six articles estimates a Cox as its main model but reported no post-estimation quantity in the main text. It did, however, report survivor curves in the appendix. To be conservative, we categorized this article as including $S(t)$.
18. We illustrate this point using a second toy example in appendix E.
19. Hanmer and Kalkan (2013) point out the way in which we select a predicted quantity's covariate values has nontrivial consequences for inference making. They contrast the incumbent "average-case approach" with the "observed-value approach" and argue in favor of the latter.
20. See appendix B for further discussion.
21. If there is overlap, we compute the confidence interval for the *first difference* of the profiles' quantity and check to see whether it overlaps with zero.

## REFERENCES

Austin, Peter C., and Janet E. Hux. 2002. "A Brief Note on Overlapping Confidence Intervals." *Journal of Vascular Surgery* 36 (1): 194–95.

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): 1260–88.

Berry, William D., Jacqueline H. R. DeMeritt, and Justin Esarey. 2010. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?" *American Journal of Political Science* 54 (1): 248–66.

Bolsen, Toby, and Judd R. Thornton. 2014. "Overlapping Confidence Intervals and Null Hypothesis Testing." *Newsletter of the APSA Experimental Section* 4 (1): 12–16.

Box-Steffensmeier, Janet M., and Bradford S. Jones. 1997. "Time Is of the Essence: Event History Models in Political Science." *American Journal of Political Science* 41 (4): 1414–61.

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

Cleves, Mario, William Gould, Roberto Gutierrez, and Yulia Marchenko. 2010. *An Introduction to Survival Analysis Using Stata*, third edition. College Station, TX: Stata Press.

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): 263–77.

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): 347–61.

Kropko, Jonathan, and Jeffrey J. Harden. Forthcoming. "Beyond the Hazard Ratio: Generating Expected Durations from the Cox Proportional Hazards Model." *British Journal of Political Science*.

Metzger, Shawna K., and Benjamin T. Jones. 2019. "Getting Time Right: Using Cox Models and Probabilities to Interpret Binary Panel Data." Working paper.

Park, Sunhee, and David J. Hendry. 2015. "Reassessing Schoenfeld Residual Tests of Proportional Hazards in Political Science Event History Analyses." *American Journal of Political Science* 59 (4): 1072–87.

Rainey, Carlisle. 2017. "Transformation-Induced Bias: Unbiased Coefficients Do Not Imply Unbiased Quantities of Interest." *Political Analysis* 25 (3): 402–409.

Schenker, Nathaniel, and Jane F. Gentleman. 2001. "On Judging the Significance of Differences by Examining the Overlap between Confidence Intervals." *The American Statistician* 55 (3): 182–86.

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