Supporting Information for “Arguing for a Negligible Effect”

TOST in R

Below, I show how you can compute the p-value for a null hypothesis of a meaningful effect using TOST. In this case, we have simple (simulated) experimental data with a treatment and control group. We hypothesize that the two groups should be similar (i.e., the treatment should have a negligible effect).

Begin by simulating data. In this case, the treatment effect is exactly zero. The treatment and control groups are coming from the exact same distributions.

```
> set.seed(9502779)
> control.group <- rnorm(100, 10, 1)
> treatment.group <- rnorm(100, 10, 1)
```

In this case, we are considering 0.5 a meaningful effect. Denoting the treatment effect Δ, our research hypothesis is $H_r : \Delta \in (-0.5, 0.5)$ and our null hypothesis is $H_0 : \Delta \in (-\infty, -0.5] \cup [0.5, \infty)$. TOST requires that we test each of the component null hypotheses and then take the largest of the two component p-values as the p-value for our overall null hypothesis. Start by testing the component null $H_0^1 : \Delta \in (-\infty, -0.5]$.

```
> c1 <- t.test(control.group, treatment.group, alt = "g", mu = -.5)
> c1$p.value
[1] 0.001220742
```

Now test the other component null $H_0^2 : \Delta \in [0.5, \infty)$.

```
> c2 <- t.test(control.group, treatment.group, alt = "l", mu = .5)
> c2$p.value
[1] 0.0002412897
```

All that is left is to find the maximum of these two p-values, which I denote as p^T.

```
> pT <- max(c1$p.value, c2$p.value)
> pT
[1] 0.001220742
```

Because we can reject both component null hypotheses $H_0^1 : \Delta \in (-\infty, -0.5]$ and $H_0^2 : \Delta \in [0.5, \infty)$, we can also reject the overall null hypothesis $H_0 : \Delta \in (-\infty, -0.5] \cup [0.5, \infty)$, by the intersection-union method (see Casella and Berger 2002, especially pp. 380-382). These data support our hypothesis of a negligible effect.

But we can also perform the exact same test by checking that the 90% confidence interval lies entirely between -0.5 and 0.5.

```
> welch <- t.test(control.group, treatment.group, conf.level = .9)
> welch$conf.int
[1] -0.2857925  0.2132902
attr(,"conf.level")
[1] 0.9
```

Here, the confidence interval is (-0.29, 0.21). Since the confidence interval lies entirely between -0.5 and 0.5, we can reject the null hypothesis of a meaningful effect. (We already knew that the confidence interval would lie between -0.5 and 0.5 because we were able to reject both component null hypotheses at the 0.05 level). However, the confidence interval contains more information, allowing us to evaluate the robustness of our decision to reject to our choice of $m = 0.5$. In this case, the confidence interval indicates that we can reject positive effects as small as 0.21 and negative effects as small as 0.29. Thus, we could have chosen m to be much smaller and still rejected the null hypothesis of a meaningful effect.

But we might like to check whether our decision to reject is also robust to our choice of statistical model. We might generate a confidence interval using the Mann-Whitney test, for example, to relax our assumption of normality in this small sample.

```
> mann.whitney <- wilcox.test(control.group, treatment.group,
+                             conf.int = TRUE, conf.level = .9)
> mann.whitney$conf.int
[1] -0.2974756  0.1764049
attr(,"conf.level")
[1] 0.9
```

The Mann-Whitney confidence interval is very similar to the confidence interval from the t-test, suggesting that our decision to reject the null hypothesis of a meaningful effect is robust to our assumption of normality.

Rather than work through the two one-sided tests "by hand," we could simply use the `tost()` function in the `equivalence` library. It relies on one-sided t-tests and exactly replicates the p^T computed above.

```
> library(equivalence)
> tost.out <- tost(control.group, treatment.group, epsilon = .5)
> tost.out$p.value
[1] 0.001220742
```

Why does $p^T = max(p^-, p^+)$?

To rigorously understand why $p^T = max(p^-, p^+)$, we first need several definitions, primarily to introduce notation, but also to precisely define several key concepts.

Types of Hypotheses

In the typical hypothesis testing situation in political science, the researcher posits a theoretically interesting hypothesis H_r, often called the alternative or research hypothesis, which suggests that some population parameter Δ, which I generally refer to as an "effect," takes on a value in a particular range. The first step in hypothesis testing requires the researcher to precisely state the theory-based prediction that she wishes to test. The research hypothesis might suggest that a variable has a non-zero, positive, negative, or negligible effect, among others.

Definition 1 (Research Hypothesis) *A research hypothesis, denoted by H_r, is a theory-based prediction that an effect of interest Δ lies in a specific region $B \subset \mathbb{R}$.*

In almost all political science applications, B is either the interval $(-\infty, 0)$, predicting a negative effect, or $(0, \infty)$, predicting a positive effect. Theories might also suggest that an effect lies in other intervals, such as $\Delta \in (m, \infty)$, where m is a threshold that defines a substantively meaningful effect.[1] In this manuscript, I argue that political scientists should consider the hypothesis that a variable does not have a substantively meaningful effect on the outcome of interest $\Delta \in (-m, m)$.

Each research hypothesis implies a particular null hypothesis. By definition, the null hypothesis suggests that $\Delta \in B^C$, or alternatively, that the true effect falls outside the interval suggested by the research hypothesis.

Definition 2 (Null Hypothesis) *Each research hypothesis implies a null hypothesis, denoted by H_0, which is that the effect lies in the region B^C.*

In this conceptual framework, the null hypothesis does not necessarily suggest that a variable has exactly no effect, although it might in some cases. This distinction is important and has key implications for the way that researchers treat hypotheses when evaluating a theory against the empirical evidence as well as implications for how researchers interpret evidence.

Testing Hypotheses

Once the researcher has carefully constructed and stated the research hypothesis and its implied null, she must evaluate the hypotheses empirically. Typically, the researcher computes a test statistic that summarizes the amount of evidence against the null hypothesis.

Definition 3 (Test Statistic) *Define a test statistic $T(\mathbf{x}) \in [0, \infty)$ as a function of the observed data $\mathbf{x}$ such that larger values of T indicate greater evidence against the null hypothesis H_0.*

Once the researcher obtains the test statistic, she must make a decision. She must decide whether she has sufficient evidence to reject the null hypothesis and therefore accept the research hypothesis as correct. If the researcher does not have enough evidence against the null hypothesis to reject it, then she fails to reject the null and the evidence is considered ambiguous. Researchers usually make this decision using a hypothesis test.

[1] Such a test would partially address the concerns raised by economists Ziliak and McCloskey (2008), who argue that statistical significance based on a null hypothesis of zero does not imply a substantively large effect.

Definition 4 (Hypothesis Test) *A hypothesis test partitions the interval $[0,\infty)$ into two regions R and R^C, such that if $T \in R$ then the analyst rejects the null hypothesis (and accepts the research hypothesis). If $T \in R^C$, then the analyst fails to reject the null hypothesis. Let $T_{crit} = \min R$, such that the analyst rejects H_0 if and only if $T(\mathbf{X}) \geq T_{crit}$. Refer to R as the rejection region and R^C as the acceptance region.*

Notice that a hypothesis test might generate two types of errors: (1) rejecting the null when it is actually correct, known as a Type-I error or false-positive, and (2) failing to reject the null when it is actually false, a Type-II error or false-negative. Obviously, we prefer tests that make fewer errors to those that make more, but we also care about the specific type of error that a test makes. In particular, convention requires that tests have a low error rate (usually less than or equal to 5%) when the true effect falls outside the region suggested by the research hypothesis (i.e., $\Delta \in B^C$). .

While political scientists sometimes report test statistics, many journal articles report other, more interpretable statistics that convey the same information. Because of this, I use p-values in the discussions below. However, the results and arguments I make depend only on using a general test statistic, not the p-value in particular.

Definition 5 (p-value) *Define a p-value such that $p = \max\limits_{\Delta \in B^C} P(T(\mathbf{X}) \geq T(\mathbf{x}))$, where $\mathbf{X}$ is a hypothetical (random) data set generated when the true effect is Δ.*

The fact that $p^T = max(p^-, p^+)$ follows directly from this definition. But to understand why, I consider three different null hypotheses: $\Delta = 0$, $\Delta \leq 0$, and $\{\Delta \leq -m \cup \Delta \geq m\}$. Political scientists are familiar with the first two cases, and once the logic of these is understood, the logic of the third follows quickly. To keep a simple running example, suppose that the parameter of interest Δ is a difference of means and the test statistic is the t-statistic.

If the researcher hypothesizes a non-zero effect, then the null hypothesis suggests that the true effect $\Delta = 0$ and B^C contains only a single point. In this situation, computing the p-value is quite easy. While $p = \max\limits_{\Delta \in B^C} P(T(\mathbf{X}) \geq T(\mathbf{x}))$ in general, we know that the maximum occurs at $\Delta = 0$ since that is the only point in the set B^C. Therefore, we can simply re-write the formula as $p = P(T(\mathbf{X}) \geq T(\mathbf{x}) | \Delta = 0)$. This reduces to the simplest hypothesis test. We simply assume that the true Δ is zero, and then calculate the probability of observing a test statistic at least as extreme as the one we actually observed.

But the logic becomes more subtle when the research hypothesis suggests that $\Delta > 0$ because B^C is now a region rather than a single point. This means we can no longer ignore the maximization across B^C. In theory, we need to show that we can reject all effects that are inconsistent with the research hypothesis. However, once we notice that the largest p-value must occur on the boundary between B and B^C ($\Delta = 0$), we can focus only on the effect suggested by the null hypothesis that is most difficult to reject. In this case, it is always $\Delta = 0$, so as before, we simply need to focus on $p = P(T(\mathbf{X}) \geq T(\mathbf{x}) | \Delta = 0)$.

Once the logic of the region becomes clear, the logic of disjoint regions immediately follows. Suppose that a researcher hypothesizes that Δ is negligible $H_r : \Delta \in (-m, m)$. Then the null hypothesis suggests that the true effect is either meaningfully negative $\Delta < -m$ or meaningfully positive $\Delta > m$. As before, we simply need to search across the space of possible effects that are consistent with the null hypothesis and argue that the largest p-value is less than 0.05. Regardless of the data, the largest p-value must again occur at the boundary between B and B^C. However, in this case, there are two boundary points, $\Delta = m$ and $\Delta = -m$. Just we only need to worry about rejecting $\Delta = 0$ in the case of the usual one-tailed test, we only need to worry about rejecting $\Delta = m$ and $\Delta = -m$ when arguing for negligible effects. We simply need to find out whether $-m$ or m generates the largest p-value. The largest p-value of the two satisfies the definition given above.

Is p^T Conservative? Why?

It turns out that p^T is theoretically conservative, although for most analyses, the impact is negligible. By conservative, I mean that if the analyst rejects H_0 if and only if $p^T \leq 0.05$, then the analyst will (incorrectly) reject the null in (strictly) less than 5% of repeated trials when the null is correct.

An important tool for evaluating and understanding tests is the power function.[2] The power function provides information about the error rate across the set of possible true effects.

Definition 6 (Power Function) *Define a power function $\beta(\Delta)$ as the probability that one rejects the null hypothesis H_0 given a true effect Δ. That is, $\beta(\Delta) = P(T(\mathbf{X}) \geq T_{crit}|\Delta)$.*

It is important to examine power functions because we expect tests to have certain properties and interpret tests as though they have these properties. The power function gives key insights into these properties. The power function also enables us to define the notion of size.

Definition 7 (Size α Test) *Say that a test is a size α test if $\max_{\Delta \in B^C} \beta(\Delta) = \alpha$. That is, the probability of rejecting the null hypothesis is less than or equal to α for all effects consistent with the null hypothesis and equal to α for at least one effect consistent with the null hypothesis.*

When I say that p^T is conservative, I mean that rejecting H_0 if and only if $p^T \leq 0.05$ does not yield a size-0.05 test in theory. However, it is very close in practice.

The reason that p^T is conservative is not immediately intuitive and best explained with the specific example below.

- I hypothesize that some effect Δ lies between -2 and 2.
- The true effect is 2. (Meaning that my research hypothesis is incorrect, but I don't know this).

[2] A power function is closely related to a power analysis. When conducting a power analysis, the researcher typically chooses a particular expected effect size that falls in the region suggested by the researcher and plugs it into the power function to obtain the probability of rejecting the null for the chosen effect. If the probability of rejecting is greater than 0.8, then the researcher claims the test to have sufficient power, since under the expected effect, the rate of false-negatives is less than 20%.

Suppose that I run the experiment many times, accepting the research hypothesis if and only if $p^T < 0.05$. We have assumed that we are using size-0.05 tests for each of the component null hypotheses, so we know that we will incorrectly reject the null hypothesis that $\Delta > 2$ in 5% of repeated trials. In some of these trials in which we erroneously reject $\Delta > m$, we fail to reject $\Delta < -m$. Thus, 5% is an upper bound for the size of the intersection-union test. If our standard errors are large relative to m, then the test might never reject the null hypothesis of no meaningful effect. However, if the test has adequate power, such that the 90% confidence interval has at most length m, then the test is very close to a size-0.05 test, since the probability of failing to reject $-m$ when the true effect is m is nearly zero.

Justifications of Robustness Checks in Reanalysis of Clark and Golder

In addition to replicating Clark and Golder's results, I present several robustness checks, each intended to be a slightly more stringent test of their hypotheses. I describe each check in Table 1. See Figure 1 in the manuscript for the estimates.

Model	Details and Motivation
Clark and Golder's Estimates	Clark and Golder use a several variables in their linear regression model, including several interactions. They choose to pool data across their main model, although they offer robustness checks that include cross-sections. They use ordinary least squares to estimate the model coefficients and Stata's cluster robust standard errors. For space concerns, I refer the reader to their article for the details of the model specification.
OLS Estimates and Standard Errors Using the Pooled Data	This model is provided purely as a reference. It makes no attempt to account for the heterogeneity across countries or the heteroskedasticity of the residuals.
Random-Intercept Model Using the Pooled Data	The random-intercept model (Steenbergen and Jones 2002; Gelman and Hill 2007) offers an alternative to clustered standard errors by modeling the homogeneity within countries.
OLS Estimate with White's Standard Errors Using Cross-Sectional Data	The (perhaps yet accounted for) within-country correlation might be driving our standard errors downward. Since this biases us toward our research hypothesis of no effect, it is important to address this concern as directly as possible. I do this by including only the most recent election from each country in the data set. I use White's standard errors (White 1980) to account for the heteroskedasticity.
M Estimator with Bootstrapped Standard Errors Using the Cross-Sectional Data.	To account for the highly non-normal errors, I use an *M* estimator (Huber 2009) that is relatively unaffected by outliers. Asymptotic standard errors seem inappropriate with $N = 49$, so I use bootstrap standard errors instead (Mooney and Duval 1993).
M Estimator with Bootstrapped Standard Errors Using the Cross-Sectional Data and Log-Transformed Outcome	While the *M* estimator is relatively unaffected by outliers, it remains susceptible to high-leverage points. One way to address some of the non-normality in the residuals is to take the natural logarithm of the outcome variable.

TABLE 1: Details and motivations for the robustness checks.

Arguments for Negligible Effects in Political Science

Despite the dearth of political methodology literature offering guidance, political scientists posit hypotheses of negligible effects quite often. I reviewed the 212 research articles published since 2011 in the *American Political Science Review* and the *American Journal of Political Science*. Of these 212 articles, 154 (73%) are empirical research articles, of which 75 (49%) present explicit hypotheses.[3] Of the 75 articles, 22 (29%) explicitly hypothesize negligible effects. (See below for a brief summary of each of these articles.) On average, more than one article per issue in the *APSR* and *AJPS* in 2011 and 2012 makes an argument that a variable should have a negligible effect.

Article	Summary
Miller (2012)	This author argues that, within dictatorships, violent leader removal makes democratization more likely. The author further argues that development (conditional on violent leader removal) has a positive effect on the likelihood of democratization but notes that "[s]ince development also makes violent leader removal less likely, the net effect of development on democratization is null (1007)."
Druckman and Leeper (2012)	These authors are interested in how pre-treatment events affect experimental treatments. They hypothesize "that pretreatment effects (e.g., leading to no experimental stimulus effect) will be more likely to occur when individuals (a) are exposed and attentive to earlier communications similar to the experimental stimuli and (b) form/update their attitudes in ways that promote strength (877-8)." They further specify that "[t]his occurs among online processors and high-NE processors (878)." Thus, these authors take a negligible treatment effect as an indication of the presence of a relevant pre-treatment event.

[3] There are several reasons why an author might not present explicit hypotheses. First, the problem might be descriptive or exploratory in nature. In this case, the author doesn't have *a priori* hypotheses. Second, the author might simply use a different style, discussing the implications of the theory in a nuanced way, rather than presenting the reader with formal hypotheses.

Heikkila and Schlager (2012)	These authors discuss the various venues used to address environmental disputes. As part of their argument, they suggest that “courts are likely to address a wide range of conflicts and are not likely to address any particular type of issue more than others (779).” Thus they suggest that the type of issue has a negligible effect on the likelihood of a court addressing the conflict.
Wright (2012)	This author hypothesizes that “[h]igh or rising unemployment will increase the Democratic vote share, regardless of the incumbent party. Lower or falling levels of unemployment will decrease the Democratic vote share, regardless of the incumbent party (692).” This, of course, implies that the incumbent’s partisanship should have a negligible impact on the effect of unemployment.
Canes-Wrone and Park (2012)	The authors present an overview of predictions from seven theories for three explanatory variables. Of the 21 total hypotheses, nine are hypotheses of a negligble effect. For example, the opportunistic political business cycles theory predicts that electoral competitiveness augments preelectoral expansion, while the traditional partisan business cycles theory predicts that electoral competitiveness should have a negligible effect. In this case, the authors use hypotheses of negligible effects to adjudicate among theoretical expectations. But the authors also use hypotheses of negligible effects as part of a complete empirical evaluation of particular theories. For example, the reverse electoral business cycles theory predicts that when policy certainty is high, the preelection period should have a negligible effect on irreversible investment.
Carrubba, Gabel, and Hankla (2012)	In a rejoinder, the authors summarize a key hypothesis from the paper under criticism: “If a government can credibly threaten noncompliance (i.e., it is a physical option), the more third-party government support and the less opposition a litigating government receives, the less retaliation for noncompliance should be anticipated, and the more likely the court is to defer to the litigating government (216).” This, of course, implies that more support and less opposition should have a negligible effect on court decisions when governments cannot credibly threaten noncompliance.

Freeman and Quinn (2012)	The authors offer a theory that suggests that financial openness and income inequality have an interactive effect on democratization. In financially open dictatorships, the authors argue, income inequality has a positive effect on democratization. This implies that in financially closed dictatorships, income inequality has a negligible effect on democratization. Further, the authors argue that greater integration into global financial markets leads to democratization in all dictatorships. But once a democratic bargain has been reached, the authors argue that neither income inequality nor global integration has a meaningful effect on democratic reversal.
Htun and Weldon (2012)	The authors observe that the United Nations Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) has been strengthening its emphasis on violence since it began in 1985. Noting that CEDAW did not emphasize violence in 1985, the authors suggest that ratification of CEDAW had little effect on the government's response to violence against women. However, they argue that in 1995 and 2005, when the language was strengthened, ratification would lead to a stronger government response.
Karpowitz, Mendelberg, and Shaker (2012)	The authors examine the effects of being in a female minority on influence within a group. Specifically, the authors argue that women's influence depends on the institution. They argue that "[i]nstitutions can eliminate the disadvantages of low numbers; similarly, they can block the power of high numbers (3)." The authors hypothesize: "Women should increase their participation with greater numbers only under majority rule; under unanimous rule, greater numbers do not benefit women, because although this rule helps minority women, it also aids minority men to the detriment of majority women (4)." Thus, the authors hypothesize that under unanimous rules, greater numbers do not meaningfully increase the influence of women in the group.

Stone Sweet and Brunell (2012)	In a response to Carrubba, Gabel, and Handkla (2008), the authors note that the first hypothesis from Carrubba, Gabel, and Handkla “implies, and rightly so, that the threat of override [of the European Court of Justice] would be credible only when a ‘sufficiently large coalition’ of member state governments (MSG) weighs in (207).” The authors claim that Carrubba, Gabel, and Handkla “treat the threat [of override] as present even in cases when only one MSG, which might be as small as Luxembourg or Portugal, has filed a brief in favor of a defendant member state (207)” and note that “this move conflicts with the precepts of intergovernmentalism, which predicts that only powerful states can constrain the court (207).” Thus, Stone Sweet and Brunell argue that support for a member state government has a negligible effect until the coalition of support becomes sufficiently large.
Weeks (2012)	This author contrasts her explanation of dictators’ decisions to initiate conflict with other theories, noting that “the conventional wisdom underestimates the vulnerability of nonpersonalist autocrats (333)” and lists several factors that “combine to produce, on average, no greater incentives for leaders of machines to initiate conflicts than for leaders of democracies (333).” Thus, in contrast with previous theories, she hypothesizes that “machines are no more likely to initiate military conflicts than democracies (333).”
Fukumoto and Horiuchi (2011)	In an article aimed at detecting voter fraud, the authors argue, for a variety of reasons, that if voters are to fraudulently register in another district, they should do so in January. Thus, they argue that holding a municipal election only increases residential registering in January. Holding a municipal election should have a negligible effect on registration in other months.

Banks and Valentino (2012)	These authors study how symbolic racism and "old-fashioned racism" moderate the effects of anger, fear, and disgust on policy attitudes. In particular, the authors hypothesize that the emotion of anger increases opposition to racially redistributive policies only among those high in symbolic racism. Thus, the authors hypothesize that anger has a negligible effect among individuals low in symbolic racism. The authors similarly hypothesize that disgust only affects policy attitudes among those high in old-fashioned racism. Of course, this leads to the hypothesis that disgust has a negligible effect among those low in old-fashioned racism. The authors also propose that fear should not meaningfully impact support for racially redistributive policies among any group.
Girod (2012)	This author examines the impact of foreign aid on development after civil wars, hypothesizing that increasing aid fosters development after civil war only when two conditions are both met: (1) when the aid recipient lacks access to rents from natural resources, and (2) when the aid is not disbursed to support donor strategic goals (e.g., military purposes). Thus, the author expects that aid has a negligible effect on development when the country has access to rents and/or the aid supports strategic donor goals.
Ono (2012)	This author examines how government portfolios are allocated within political parties. In particular, he discusses how internal strength (popularity within the party) and external popularity (with the public) interact in their influence on the seat shares assigned to the leader's faction. The author argues that external popularity should lead to greater portfolio shares for the leader's faction if the party leader leads an internally weak faction. On the other hand, if the party leader leads an internally strong faction, then external popularity should have a negligible effect on portfolio shares. Similarly, if the leader is externally popular, then internal strength should have a negligible effect on portfolio shares.

Pepinsky, Liddle, and Mujani (2012)	These authors present several possible explanations of why Islamic parties seem to have an advantage in the Muslim world. The authors enumerate the predictions from each of these possible explanations in their Table 1. For example, the authors' preferred theory implies that being an Islamic party should have no effect on support when voters have favorable or unfavorable views toward the party's economic platform. Only when voters are uncertain about the party's economic platform, the authors argue, does its Islamic status increase support. Therefore, the authors suggest that when voters have favorable or unfavorable views toward a party's economic platform, its Islamic status should have a negligible effect on support. As a contrast, one competing explanation implies that Islamic status should always increase support, regardless of the individuals' views toward the economic platform of the party.
Gibler and Randazzo (2011)	These authors examine whether independent judiciaries can increase democratic survival. They argue, however, that only established judiciaries are able to accomplish this. New judiciaries, they argue, have a negligible effect on the probability of democratic survival, since new judiciaries tend to reflect the interest of elites.
Greene (2011)	This author argues that, especially in new democracies, campaigns can persuade voters. However, he argues that campaign effects occur only among voters with weak partisan attachment and less political knowledge. Campaigns, the author argues, have a negligible impact on voters with strong partisan attachments and high political knowledge.
Krupnikov (2011)	This author examines the conditions under which negative advertising influences political participation. Specifically, she argues that late negativity targeted toward a liked candidate has a demobilizing effect. However, she argues that negativity directed toward a disliked candidate should have a negligible effect on participation.

Ryan (2011)	This author examines the influence of information from various sources on voters' decision making. In particular, the author hypothesizes that information from out-party sources will have a negligible effect on individuals' vote choice. The author also argues that poorly-informed individuals benefit from information from other individuals with the same partisanship. Therefore, information from out-partisans has only a negligible effect on individuals' probability of voting correctly.

References

Banks, Antoine J., and Nicholas A. Valentino. 2012. "Emotional Substrates of White Racial Attitudes." *American Journal of Political Science* 56(2):286–297.

Canes-Wrone, Brandice, and Jee-Kwang Park. 2012. "Electoral Business Cycles in OECD Countries." *American Political Science Review* 106(1):103–122.

Carrubba, Clifford J., Matthew Gabel, and Charles Hankla. 2012. "Understanding the Role of the European Court of Justice in European Integration." *American Political Science Review* 106(1):214–223.

Carrubba, Clifford, Matthew Gabel, and Charles Handkla. 2008. "Judicial Behavior Under Political Constraints: Evidence from the European Court of Justice." *American Political Science Review* 102(4):435–452.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. 2nd ed. Pacific Grove, California: Duxbury.

Druckman, James N., and Thomas J. Leeper. 2012. "Learning More From Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56(4):875–896.

Freeman, John R., and Dennis P. Quinn. 2012. "The Economic Origins of Democracy Reconsidered." *American Political Science Review* 106(1):58–80.

Fukumoto, Kentaro, and Yusaku Horiuchi. 2011. "Making Outsiders' Votes Count: Detecting Electoral Fraud through a Natural Experiment." *American Political Science Review* 105(3):586–603.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gibler, Douglas M., and Kirk A. Randazzo. 2011. "Testing the Effects of Independent Judiciaries on the Likelihood of Democratic Backsliding." *American Journal of Political Science* 55(3):696–709.

Girod, Desha M. 2012. "Effective Foreign Aid Following Civil War: The Nonstrategic-Desperation Hypothesis." *American Journal of Political Science* 56(1):188–201.

Greene, Kenneth F. 2011. "Campaign Persuasion and Nascent Partisanship in Mexico's New Democracy." *American Journal of Political Science* 55(2):398–416.

Heikkila, Tanya, and Edella C. Schlager. 2012. "Addressing the Issues: The Choice of Environmental Conflict-Resolution Venues in the United States." *American Journal of Political Science* 56(4):774–786.

Htun, Mala, and S. Laurel Weldon. 2012. "The Civic Origins of Progressive Policy Change: Combating Violence against Women in Global Perspective, 1975–2005." *American Political Science Review* 106(3):1–22.

Huber, Peter J. 2009. *Robust Statistics*. Wiley.

Karpowitz, Christopher F., Tali Mendelberg, and Lee Shaker. 2012. "Gender Inequality in Deliberative Participation." *American Political Science Review* 106(3):1–15.

Krupnikov, Yanna. 2011. "When Does Negativity Demobilize? Tracing the Conditional Effect of Negative Campaigning on Voter Turnout." *American Journal of Political Science* 55(4):797–813.

Miller, Michael K. 2012. "Economic Development, Violent Leader Removal, and Democratization." *American Journal of Political Science* 56(4):1002–1020.

Mooney, Christopher Z., and Robert D. Duval. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Sage University Press.

Ono, Yoshikuni. 2012. "Portfolio Allocation as Leadership Strategy: Intraparty Bargaining in Japan." *American Journal of Political Science* 56(3):553–567.

Pepinsky, Thomas B., R. William Liddle, and Saiful Mujani. 2012. "Testing Islam's Political Advantage: Evidence from Indonesia." *American Journal of Political Science* 56(3):584–600.

Ryan, John Barry. 2011. "Social Networks as a Shortcut to Correct Voting." *American Journal of Political Science* 55(4):753–766.

Steenbergen, Marco R., and Bradford S. Jones. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46(1):218–237.

Stone Sweet, Alec, and Thomas Brunell. 2012. "The European Court of Justice, State Noncompliance, and the Politics of Override." *American Political Science Review* 106(1):204–213.

Weeks, Jessica L. 2012. "Strongmen and Straw Men: Authoritarian Regimes and the Initiation of International Conflict." *American Political Science Review* 106(2):326–347.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4):817–838.

Wright, John R. 2012. "Unemployment and Democratic Electoral Advantage." *American Political Science Review* 106(4):685–702.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.