Meaningful Inferences

Using Confidence Intervals to Account for Uncertainty in Substantive Interpretations*

Kelly McCaskey†

Carlisle Rainey‡

Abstract

Research in political science is gradually moving away from an exclusive focus on statistical significance testing and toward an emphasis on effect magnitude. We argue that the current practice of "magnitude-and-significance," in which researchers interpret only the magnitude of a statistically significant point estimate is only a small improvement over the much maligned "sign-and-significance" approach, in which researchers focus only on the statistical significance of an estimate. Instead of interpreting the point estimate alone, researchers should explicitly account for uncertainty when interpreting effect sizes. We suggest that researchers interpret the range of values contained in the confidence interval and avoid making claims about the point estimates that do not hold for the entire confidence interval. Using the effect of U.N. troops on civilian casualties during civil war, we demonstrate how this approach clarifies statistical arguments by better communicating the uncertainty of the estimates.

> Far better an approximate answer to the *right* question, which is often vague, than an exact answer to the *wrong* question, which can always be made precise.
>
> Tukey (1962, pp. 13-14)

Manuscript word count: 3,995

*We thank Lisa Hultman, Jacob Kathman, and Megan Shannon and Cindy Kam and Elizabeth Zechmeister for making their data available to us. The analyses presented here were conducted with R 3.1.0. All data and computer code necessary for replication are available at github.com/carlislerainey/meaningful-inferences .

†Kelly McCaskey is a Ph.D. student in the Department of Political Science, University at Buffalo, SUNY, 520 Park Hall, Buffalo, NY 14260 (kellymcc@buffalo.edu).

‡Carlisle Rainey is Assistant Professor of Political Science, University at Buffalo, SUNY, 520 Park Hall, Buffalo, NY 14260 (rcrainey@buffalo.edu).

Introduction

Recent work in political science encourages researchers to move beyond statistical significance and focus on the *substantive importance* of the *magnitude* of the effect (e.g., King, Tomz, and Wittenberg 2000; Hanmer and Kalkan 2013; and Esarey and Danneman 2014). The current practice in political science argues for substantive importance by interpreting the magnitude of statistically significant point estimates. If the estimate is substantively large, then the researcher concludes the effect is "substantively and statistically significant." However, this approach is not compelling. Indeed, we show that focusing *only* on the magnitude of the estimated effect retains many of the flaws associated with focusing *only* on statistical significance. Instead, researchers should take the uncertainty of the estimates into account when arguing for substantive significance by interpreting the range of values in the confidence interval, not just the point estimate. This idea is not new (e.g., Achen 1982, Gross 2014, and Rainey 2014), but it has not become common practice. Our aims are twofold. First, we explain the problem with the current practice, and, second, we explain how researchers can use confidence intervals to more transparently evaluate substantive claims. Our key point is that researchers should interpret the range of values contained in their confidence intervals, and, conversely, avoid making claims that are not consistent with the range of values in their confidence intervals.

We contend that the best empirical political science concerns itself with the (1) direction, (2) size, and (3) substantive importance of the effects of interest. Each of these questions requires increasing levels of analysis and substantive interpretation and each receives decreasing levels of attention in political science research. Virtually every research article in empirical political science makes an argument about the direction of the effect using a null hypothesis significance test. Almost all of these articles report some measure of effect size, even if it is just in the form of a table of regression coefficients.

However, judgments about the importance of the effect are often missing. To get a sense of how often researchers address substantive importance (i.e., Are the effects large enough to care about?), we reviewed all articles published in the *American Political Science Review* and the *American Journal of Political Science* from 2011 to 2013.[1] Of the 316 total articles, 73% were empirical analyses. Of this 73%, only about half make a judgment about the substantive importance of the estimated effect. Furthermore, of the articles that discuss the substantive importance of their results, only 17% make an explicit argument that the estimates are substantively large and discuss these implications.

We begin by elaborating on the questions that political scientists typically ask about effects and explain the typical approaches to answering these questions. We then provide an overview of the much maligned sign-and-significance approach and the current best practice of magnitude-and-significance. Additionally, we explain why the magnitude-and-significance approach is only a small improvement over the sign-and-significance approach. As an alternative, we suggest that researchers focus on the range of values contained in the confidence interval and avoid focusing exclusively on the point estimate. We conclude an empirical example that highlights the importance of substantively interpreting the range of effects that are consistent with the data.

What We Want To Know

Most empirical research in political science focuses on estimating the effect of an explanatory variable on the expected value of an outcome of interest.[2] Importantly, empirical work usually

[1] To do this, we read the 316 articles published in the range of years and coded whether or not it was an empirical study. Of those that were empirical, if the author made the claim that the estimate was large because of some theory-based reason, they were coded as having made an explicit argument about the substantive nature of their results.

[2] Formally, denote the outcome variable y and the explanatory variable x. Suppose further that $E(y|x) = f(x)$ and define the "effect" or "quantity of interest" Δ as the difference between the average outcome when x takes on a substantively meaningful high value and low value, so that $\Delta = E(y|x = x_{hi}) - E(y|x = x_{lo}) = f(x_{hi}) - f(x_{lo})$. We adopt this notation to be consistent with the situation in which researchers are focusing on the first-difference

focuses on answering three fundamental questions about the effect of the explanatory variable on the outcome.

1. What is the direction of the effect?
2. How large is the effect?
3. Is the effect substantively important?

Direction

The first question that empirical research usually attempts to answer is the direction of the effect. Is the effect positive or negative?[3]

Suppose, for clarity, that the researcher offers a directional research hypothesis, suggesting that an effect of interest is positive.[4] Then the researcher would compare this research hypothesis to a null hypothesis that suggests that the research hypothesis is false, or equivalently, that the effect lies outside the region suggested by the research hypothesis.[5] To assess the evidence against the null hypothesis, the researcher would usually calculate a p-value, which is the probability of obtaining (hypothetical) data at least as extreme as the observed data if the null hypothesis were true. If this p-value is sufficiently small (by convention, less than 0.05), then the researcher would reject the null hypothesis in favor of the research hypothesis or, alternatively, reject negative effects and conclude that the effect is positive. However, if the p-value is not sufficiently small, then the researcher would declare that the data do not offer compelling evidence against the null hypothesis and note that the direction of the effect remains uncertain.[6]

as their key quantity of interest (King, Tomz, and Wittenberg 2000).

[3]Some research argues theoretically and empirically for "no effect" or "a negligible effect" (e.g., Kam and Palmer 2008, see Rainey 2014), but most hypotheses posit the direction of an effect.

[4]Formally, we might denote this hypothesis as $H_r : \Delta > 0$.

[5]Formally, we might write this as $H_0 : \Delta \leq 0$.

[6]However, some research takes a p-value greater than 0.05 as evidence *in favor of* the null hypothesis (Rainey 2014). We prefer to interpret a lack of statistical significance as ambiguous evidence from which the researcher can make no claim (i.e., the effect might be negative or positive).

Magnitude

However, recent methodological work points out that empirical research should go beyond estimating the direction of the effect and emphasize the size or magnitude of the effect as well (King, Tomz, and Wittenberg 2000; Hanmer and Kalkan 2013; Gross 2014).[7] In discussing how scholars might interpret a model of the effects of education on income, King, Tomz, and Wittenberg (2000, p. 348) write:

> Bad interpretations are substantively ambiguous and filled with methodological jargon: "the coefficient on education was statistically significant at the 0.05 level." Descriptions like this are very common in social science, but students, public officials, and scholars should not need to understand phrases like "coefficient," "statistically significant," and "the 0.05 level" to learn from the research. Moreover, even statistically savvy readers should complain that the sentences does not convey the key quantity of interest: how much higher the starting salary would be if the student attended college for an extra year.

The emphasis on effect magnitude is hardly a new idea. Commenting on the consequences of Fisher's null hypothesis significance test, Yates (1951, p. 32) writes: "[I]t has caused scientific workers to pay undue attention to the results of the tests of significance they perform on their data, particularly data derived from experiments, and too little to the estimates of the magnitude of the effects they are investigating."

Fortunately, recent conceptual work (King, Tomz, and Wittenberg 2000; Berry, DeMeritt, and Esarey 2010; and Hanmer and Kalkan 2013) and software development (Tomz, Wittenberg,

[7] King, Tomz, and Wittenberg (2000) show that computing an interpretable measure of effect magnitude is not always straightforward. Only some statistical models have naturally interpretable parameters. For example, a simple difference-in-means or normal-linear model has directly interpretable coefficients as long as the scales of the variables are reasonable and the model does not include non-linear or product terms. Outside of this atypical situation, however, the researcher must do additional work to estimate a substantively meaningful quantity of interest.

and King 2003; and Imai, King, and Lau 2008) empower political scientists to move beyond a simple "sign-and-significance" approach and also present substantively meaningful measures of effect magnitude.

Importance

But researchers ultimately want to move beyond the simple presentation of effect magnitude and make a judgment about the meaningfulness of the effects. Is the effect large or small? Is it substantively important? Is it relevant for policy? Is it scientifically important? Is it large enough to matter? Hanmer and Kalkan (2013, p. 264) write: "[W]e take it as given that understanding whether the relationship is substantively significant, rather than just statistically significant, is the ultimate goal, as it is a necessary part of vaulting one's theory."

As Rainey (2014) notes, political scientists should not insist on hard and fast rules for judging the effects that are and are not substantively meaningful.[8] Instead, we must insist that substantive scholars making substantive claims about politics also make substantive judgments about the importance of their effects. Thompson (2001, pp. 82-83) notes for example, that "if people interpreted effect sizes with the same rigidity that $\alpha = 0.05$ has been used in statistical testing, we would merely be being stupid in another metric." Kirk (1996) notes that this judgment is "influenced by a variety of factors, including the researcher's value system, societal concerns, assessment of costs and benefits, and so on." Despite this element of subjectivity, Thompson (2002, p. 30) writes:

> The existence of effect size benchmarks should not justify abrogating the responsibility for arguing for effect import in the specific context of a given study. It is not necessary to have universal benchmarks regarding what effect sizes may be deemed noteworthy. The reader with a value system widely different than that of

[8] We should note, though, that such rules of thumb have been presented, see Glass (1976) and Cohen (1992), but these rules are usually proposed with caution.

> an author might reasonably disagree with the author about whether the effect size is noteworthy and then simply ignore the study.

Substantive judgments about effect sizes require a large initial investment of careful thought and this judgment demands subjectivity (see Rainey 2014). However, because this subjective judgment is transparent, readers are free to reject the author's judgment and substitute their own. Further, "automatic" and "objective" procedures are not always (or perhaps usually) desirable.[9] Substantive scholars making substantive points about politics must not be prohibited from making substantive judgments. Instead, they must be *encouraged* to do so (Achen 1982). Indeed, Kirk (1996, p. 755) writes:

> Researchers have an obligation to make this kind of judgment. No one is in a better position than the researcher who collected and analyzed the data to decide whether or not the results are trivial. It is a curious anomaly that researchers are trusted to make a variety of complex decisions in the design and execution of an experiment, but in the name of objectivity, they are not expected or even encouraged to decide whether data are practically significant.

For example, King and Zeng (2001, p. 711) claim that, "if a collection of 300,000 dyads shows a 0.001 increase in the probability of war, the finding is catastrophically *important because* it represents about three hundred additional wars and a massive loss of human life" (emphasis ours). In another example, Hetherington and Suhay (2011, p. 317) note that

> When we fixed trust at its minimum, the predicted probability that our typical respondent thought Iraq was worth the cost was only 0.33; we would classify

[9] Formal hypothesis tests and judgments about substantive importance are qualitatively different decisions and have different strengths and weaknesses. Estimation and hypothesis tests are relatively automatic and "objective," but not transparent. Researchers do not fit one model and report the single p-value. Instead they fit many models and report the one that "makes most sense" in light of their approach, theoretical model, normative concerns, and the results of the model. (Gerber and Malhotra 2008; Simmons, Nelson, and Simonsohn 2011; Francis 2013; Simmons, Nelson, and Simonsohn 2014; and Esarey and Wu 2014; see also Gelman and Loken 2014).

him as not believing Iraq was worth the cost. If we increase political trust to its maximum, however, the predicted probability more than doubles to 0.72, a 39 percentage point increase. This is *substantively important because*, with trust at its maximum, our typical respondent believes Iraq was worth the cost.

This raises three questions about the process of judging an effect to be important or unimportant.

1. What is the current practice for making these judgements?
2. Is this practice compelling?
3. If not, what might serve as a better approach?

We now turn to these questions.

Current Practices in Reasoning About Effect Importance

Sign-and-Significance

Using the sign-and-significance approach, researchers declare an effect important if and only if the effect is statistically significant. That is, researchers simply test whether the effect is greater than (or less than) zero. If they reject the null (e.g., $p < 0.05$), then they declare, or perhaps subtly imply, the effect to be substantively meaningful. While this approach is not commonly employed in political science research, it is important to highlight because its shortcomings are commonly understood (Gill 1999).

The much maligned (e.g., Cohen 1990, Gill 1999, Hill and Jones 2014, and Gross 2014) sign-and-significance approach emphasizes p-values, but a minuscule p-value does not imply a large or important effect. The p-value depends on both the sample size and the effect size. It is true that the p-value gets smaller as the effect size under investigation increases. However, the p-value also decreases as the sample size increases, which is unrelated to the effect size. While it is common to describe estimates as "highly significant" or "very significant," this implies

(or tempts readers to conclude) that an effect is large or important. "Very significant" (e.g., $p < 0.001$) might simply mean that the researcher has a large sample. Even if the researcher finds statistically significant results with a small sample, the small p-value only indicates that the effect size is large relative to the uncertainty. It says nothing about the size of the estimate relative to some standard of substantive importance. With a very small p-value (e.g., $p < 0.001$), substantive experts might judge the effect to be large, moderate, small, or even negligible. Experts might call some of these effects important, somewhat important, slightly important, or not at all important. The p-value is indirectly related to substantive significance at best.

Magnitude-and-Significance

The standard scientific practice in political science is to take the magnitude-and-significance approach, in which the researcher:

1. Computes substantively interpretable estimate of the effect of interest, such as a first difference.
2. Tests whether these estimates are statistically significant.
3. If so, makes a judgement about whether the magnitude is substantively important.

For example, Tomz and Weeks (2013) discuss the results of their experiment on attitudes about conflict.

> Citizens in both countries were much less willing to attack another democracy than to attack an otherwise equivalent autocracy. Approximately 34.2% of respondents in the U.K. supported a military strike when the country was not a democracy versus 20.9% when the country was a democracy. Thus, democracy reduced support for a military strike by more than 13 percentage points, with a 95% confidence interval of 19.6 to 6.9. The baseline level of militarism was much higher in the United States, where at least half the respondents wanted to strike an autocracy. Nonetheless,

> democracy exerted a similarly large effect in the United States: The between-subjects and within-subject estimates concurred that democracy reduced enthusiasm for a military strike by about 11.5 percentage points. In both countries, democracy produced *substantively large and statistically significant* effects on preferences (p. 854-855, emphasis ours).

Gerber and Hopkins (2011) discuss their regression model.

> The first model indicates that all else equal, a city where the Democrat just wins the mayoralty should expect its spending on police to drop by 2.3 percentage points three fiscal years later. This result is *statistically significant*, with a 95% confidence interval that runs from 0.5 percentage points to 4.0 percentage points. It is *substantively large as well*, as it reflects a spending shift of 1.2 standard deviations in terms of the dependent variable (p. 333, emphasis ours).

Ansolabhere and Jones (2010) discuss the coefficients of their regression model.

> The regression results in [their] Table 3 show that a Representative's actual roll-call votes strongly predict respondents' beliefs about the Representative's votes. The coefficients for both 2005 and 2006 are *substantively large* (approximately 0.3) *and statistically significant* (p. 588, emphasis ours).

The magnitude-and-significance approach has three unfortunate consequences. It tempts researchers to...

1. treat all statistically significant and substantively large estimates similarly, drawing no distinction between large, imprecise estimates and large, precise estimates.
2. treat "barely significant" large estimates and "almost significant" large estimates quite differently, drawing a strong distinction between two similar estimates with similar uncertainty.

3. treat all results that are not statistically significant similarly, drawing no distinction between large, imprecise estimates and small, precise estimates (see Rainey 2014).

Consider, for example the hypothetical studies presented in Figure 1. If we are interested in substantive significance and consider 0.25 as the smallest substantively meaningful effect, then only one study, Study A, seems to offer strong evidence for a substantively meaningful effect. After that, Studies B and C offer ambiguous evidence for a meaningful effect and Studies D and E offer evidence *against* a meaningful effect.

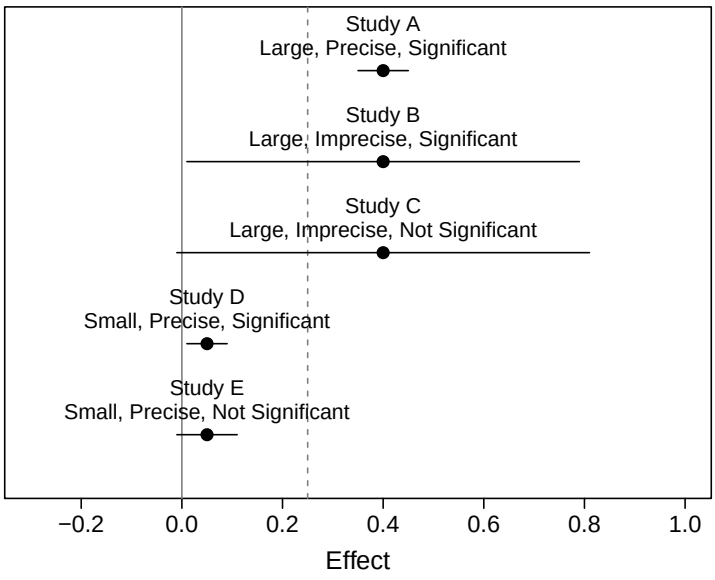

FIGURE 1: This figure provides several hypothetical studies to illustrate several points about arguments for substantive significance. Notice that Study A offers compelling evidence for a meaningful effect, Studies B and C are consistent with both small and large effects, and Studies D and E offer evidence *against* meaningful effects. However, the current practice in political science of magnitude and significance treats Studies A and B similarly, Studies B and C *differently*, and Studies C and E similarly. Table 1 compares the interpretations from both the sign-and-significance and magnitude-and-significance approaches with the intuitive meaning.

Although only Study A offers compelling evidence for a meaningful effect, the magnitude-and-significance approach suggests that Studies A and B offer similar evidence for a substantively meaningful effect as well, since both are "statistically significant and substantively large."

Yet these studies do not offer similar evidence for a meaningful effect. Study A is only consistent with a meaningful effect. Study B, on the other hand, is also consistent with small, negligible effects.

In fact, the amount of evidence for a meaningful effect offered by Study B is similar to that offered by Study C–both studies are consistent with both large and small effects. Yet the magnitude-and-significance approach treats these studies *differently*. The magnitude-and-significance approach concludes from Study B that the effect is "positive, significant, and substantively meaningful" and from Study C that the effect is "not statistically significant."

The magnitude-and-significance approach also treats Studies C and E similarly because both estimates are "not statistically significant." However, Study C is consistent with large *and* small effects. Study E on the other hand, is consistent with *only* small effects.

Indeed, the only improvement of the magnitude-and-significance approach over the much maligned sign-and-significance approach is that the magnitude-and-significance approach manages to distinguish between Study A and Study D. Other than that, both of these approaches treat Study A and B similarly, Studies B and C differently, and Studies C and E similarly. From our perspective, these are all inferential errors. Table 1 shows the interpretation from each approach compared with an "intuitive approach" based on the confidence intervals. Especially notice the similarity of the errors made by the dismissed sign-and-significance approach and the current practice of magnitude-and-significance.

Study	Sign-and-Significance Method	Magnitude-and-Significance	Intuitive Interpretation
Study A	"positive and significant"	"positive, significant, and substantively large"	"We have strong evidence for a large, substantively meaningful effect."
Study B	"positive and significant"	"positive, significant, and substantively large"	"We have only weak evidence for a large, substantively meaningful effect, because the data are also consistent with negligible effects near zero."
Study C	"not statistically significant"	"not statistically significant"	"We have only weak evidence for a large, substantively meaningful effect, because the data are also consistent with negligible effects near zero."
Study D	"positive and significant"	"positive and significant, but substantively small'	"We have strong evidence against a substantively meaningful effect."
Study E	"not statistically significant"	"not statistically significant"	"We have strong evidence against a substantively meaningful effect."

TABLE 1: This table compares the interpretations of the results in Figure 1 using the sign-and-significance and the magnitude-and-significance approaches to the intuitive meaning of the confidence intervals. Notice that the much maligned sign-and-significance approach and the magnitude-and-significant approach only differ in their interpretations of Studies D and E. Both approaches deviate substantially from the intuitive interpretation based on the confidence intervals.

It is important to apply similar standards of evidence to arguments for positive (or negative) effects and arguments for *meaningfully* positive (or negative) effects. The usual logic of hypothesis testing requires that the researcher only declare that an effect is positive (or negative) if and only if the evidence overwhelmingly points toward a positive (or negative) effect. Similarly, researchers should not declare a positive estimate to be substantively meaningful simply because it is inconsistent with negative effects and the point estimate lies above some threshold. A consistent standard of evidence requires that researchers declare an effect to be substantively meaningful if and only if it is inconsistent with negligible effects.

Confidence Intervals

As one solution to the problem, Gross (2014) presents the formal PASS-test. Using the PASS-test, researchers specify a pre-chosen value (that we denote as) m that represents the smallest substantively interesting effect. Values larger than m are thought to be substantively important and values smaller than m are thought to be substantively *un*important.

While this formal hypothesis testing framework is sometimes clear and convenient, confidence intervals offer even more information and are easier for researchers (and readers) to interpret. Importantly, there is a one-to-one correspondence between a hypothesis test and a confidence interval. Specifically, a 90% confidence interval contains only values greater than m if and only if a size-0.05 hypothesis test rejects the null hypothesis that the effect is less than or equal to m. Therefore, if the 90% confidence interval contains only large, meaningful effects, then the researcher can confidently reject the null hypothesis of a small, negligible effect. However, if the 90% confidence interval contains effects that are inconsistent with the hypothesis of a meaningful effect, such as small, negligible effects, the evidence for the researchers claim is (correctly) identified as weaker.[10]

But do researchers need to specify a pre-chosen m in *practice*? Choosing a specific m is certainly useful to discuss formal tests for meaningful effects, but Tukey (1991, pp. 101-102) warns researchers about phony precision.

> The precise logic of mathematics serves statistician and data analyst in derivations–in theoretical structures which do help us in thinking about the world. But how we think about the world needs to be suitably imprecise. We dare not limit ourselves to such formal precision.

Thus, in practice, we suggest that researchers avoid choosing an arbitrary cut point and focus instead on interpreting the range of effects consistent with the data. While specifying a pre-chosen m is useful in setting up the theoretical argument, such a pre-chosen m is artificially precise for social science practice and does not sufficiently acknowledge the continuum (as

[10] Formally, $100(1-\alpha)\%$ confidence interval contains the set of values that cannot be rejected by a size-α two-tailed test. Thus, all values $u_{\alpha}^{+/-}$ that fall outside (i.e., above or below) the confidence interval are rejected by a two-tailed test of size α. Confidence intervals have a similar relationship with one-tailed tests. All values $u_{2\alpha}^{-}$ that fall *below* a $100(1-2\alpha)\%$ are rejected by a one-tailed test of the null hypothesis that the true parameter lies at or below $u_{2\alpha}^{-}$. Similarly, all values $u_{2\alpha}^{+}$ that fall *above* a $100(1-2\alpha)\%$ are rejected by a one-tailed test of the null hypothesis that the true parameter lies at or above $u_{2\alpha}^{+}$. Thus, there is a one-to-one correspondence between one- and two-tailed hypothesis tests of size 0.05 and 90% and 95% confidence intervals respectively (see esp. Casella and Berger 2002, pp. 419-423).

opposed to a cutpoint) between meaningful and negligible effects.

Rather than pre-specifying m, we suggest that researchers take a "softer" approach and compute the confidence intervals for the substantively interpretable effects first. With the estimates and confidence intervals in hand, we suggest that researchers then interpret the range of effects contained in the confidence interval and only interpret the estimate as "substantively significant" if the confidence interval contains *only* substantively meaningful values.[11] In essence, we recommend following Achen's (1982) advice.

> What general advice can be given for interpreting confidence intervals? The best use of them depends on the problem at hand, and no universal instructions can be given. However, one rarely errs by giving a 95% interval, explaining what the endpoints would mean substantively if each were true, and interpreting the overall results in such a way as to allow for the possibility that either of those endpoints is, in fact, the truth (p. 50).

Replication of Hultman, Kathman, and Shannon (2013)

To illustrate how this idea might work in practice, we now turn to a study by Hultman, Kathman, and Shannon (2013).[12] The authors explain that civilians can be successfully protected by UN peacekeeping operations (PKOs) when those missions are composed of military troops and police in adequately large numbers. They argue that PKOs mitigate violence both on the battlefield and behind the battlefield's frontlines for a variety of reasons and that the UN's ability to intervene is contingent upon the size and personnel composition of the deployment.

Specifically, Hultman, Kathman, and Shannon hypothesize that as the UN commits both more military troops to a conflict, the amount of violence committed against civilians will

[11] In Appendix A, we note several potential situations in which the researcher might want to avoid focusing on the magnitude of the effect and instead focus in the direction

[12] In Appendix B, we discuss another study by Kam and Zechmeister (2013).

decrease. These authors argue for a meaningful effect by (1) showing that the relevant quantity of interest is statistically significant and then (2) suggesting that the estimated effect is substantively meaningful.

Following the advice of the literature on interpreting the magnitude of the effects Hultman, Kathman, and Shannon present a plots of the expected civilian deaths as the number of military and policy troops varies. We replicate this plot in Figure 2.

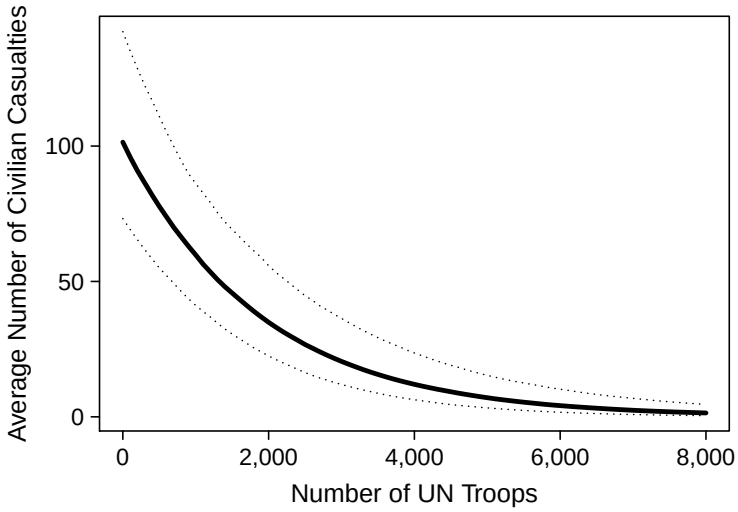

FIGURE 2: This figure shows expected number of civilian casualties as the number of U.N. troops increases.

Noting that the relevant coefficients are statistically significant and correctly signed, Hultman, Kathman, and Shannon write:

> The negative and statistically significant ($p < 0.001$) effects of *UN Military Troops* and *UN Police* suggest that as PKOs are increasingly supplied with soldiers and police forces, violence against civilians in civil war decreases (p. 9-10).

The authors argue that effect is not only statistically significant, but substantively important.

> The figure shows that increasing the number of troops has a *dramatic effect* on improving the safety of noncombatants. With no troops deployed to a conflict,

> the expected number of civilians killed in a given month is approximately 106. When the number of UN military troops increases to 8,000, the expected value of civilian deaths declines to 1.79. Conditional on the other variables being held at the specified values, supplying only several thousand military troops *nearly mutes* violence completely as the number of troops approaches the upper values reported (p. 11, emphasis ours).

They continue:

> Bear in mind that the values presented are expected civilian deaths per month. These are *not inconsequential* reductions in violence. Indeed, given that the average length of a conflict in these data is nearly 65 months, deploying highly equipped missions can *mitigate or wholly avert humanitarian disasters* (p. 11, emphasis ours).

However, they do not explicitly take uncertainty into account when arguing for a meaningful effect. Instead, they only check that the *point estimate* is substantively important. They do not consider whether trivial effects are also plausible based on the data.

To assess whether their substantive claim is robust to accounting for the uncertainty, we replicate their results and calculate the expected changes in civilian deaths as UN military troops increases and 90% confidence intervals. Figure 3 shows these confidence intervals. At an expense of roughly $2 million, 2,000 troops leads to about 65 fewer civilian casualties, on average. However, the data suggest that the effect leads to *at least* 45 fewer civilian casualties and possibly as many as 95. Similarly, an expense of about $8 million, or 8,000 troops, leads to the prevention of approximately 100 civilian casualties. However, we can be confident that 8,000 troops prevents *at least* 70 civilian deaths and perhaps as many as 140.

In this case, the authors indeed have strong evidence for a *dramatic effect*, even after uncertainty is taken into account. While the data support their claim, the authors can make a

stronger argument for a meaningful effect by explicitly accounting for uncertainty by substantively interpreting the range of plausible values rather than the point estimate.

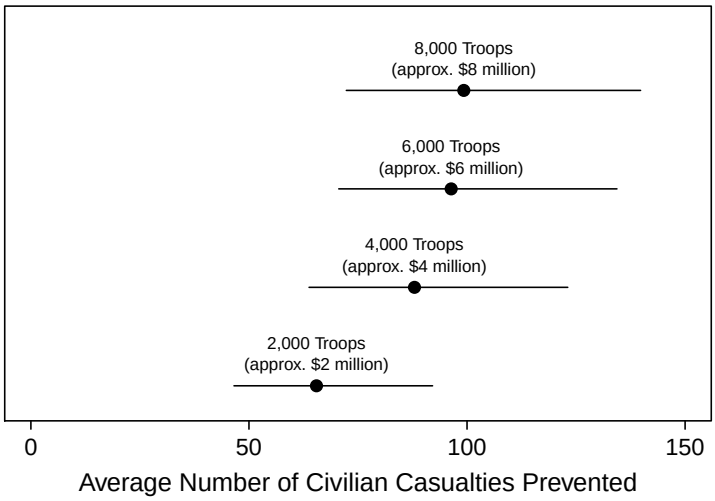

FIGURE 3: This figure provides first differences, in dollar and troop amounts, and their corresponding number of prevented civilian casualties.

Conclusion

In this paper, we have argued that researchers should account for uncertainty when substantively interpreting their statistical results. Our suggestion that researchers formally test that effects lie above a threshold of substantive significance is not new (Achen 1982, Rainey 2014, and Gross 2014; see also Esarey and Danneman 2014). However, explicit testing of substantive claims is not yet common practice and scholars rarely offer complete, substantive interpretations of the range of effects contained in confidence intervals. The current practice continues to be testing a directional research hypothesis and interpreting the substantive meaning of the estimate without taking into account the uncertainty surrounding the estimate.

We hope that our discussion encourages researchers to move beyond the current practice and adopt two conventions. First, we hope that researchers will begin to interpret the range

of values contained in the confidence intervals. Second, we hope that researchers will avoid making substantive claims based on point estimates when these claims are not also consistent with the range of values contained in the confidence intervals. Using this approach, researchers will (1) compute quantities that are of direct substantive interest, (2) clarify claims about the effects they consider theoretically and/or normatively important, and (3) take the uncertainly of the estimates into account when assessing the evidence for their substantive claims. This leads more transparent substantive claims and clearer communication of the empirical evidence for these claims.

References

Achen, Christopher H. 1982. *Interpreting and Using Regression*. Thousand Oaks, CA: Sage.

Ansolabhere, Stephen, and Philip Edward Jones. 2010. "Constituents' Responses to Congressional Roll-Call Voting." *American Journal of Political Science* 54(3):583–597.

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):105–119.

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

Cohen, Jacob. 1990. "Things I Have Learned (So Far)." *American Psychologist* 45(12):1304–1312.

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112(1):115–159.

Esarey, Justin, and Ahra Wu. 2014. "The Fault in our Stars: Measuring and Correcting Significance Bias in Political Science." Working paper. Copy at jee3.web.rice.edu/significance-bias.pdf.

Esarey, Justin, and Nathan Danneman. 2014. "A Quantitative Method for Substantive Robustness Assessment." Forthcoming in *Political Science Research and Methods*.

Francis, Gregory. 2013. "Replication, Statistical Consistency, and Publication Bias (with Discussion)." *Journal of Mathematical Psychology* 57(5):153–169.

Gaines, Brian J., and James H. Kuklinski. 2011. "Experimental Estimation of Hetergeneous Treatment Effects Related to Self-Selection." *American Journal of Political Science* 55(3):724–736.

Gelman, Andrew, and Eric Loken. 2014. "Ethics and Statistics: The AAA Tranche of Subprime Science." *CHANCE* 27(1):51–56.

Gerber, Alan. 2011. "Field Experiments in Political Science." In *Handbook of Experimental Political Science*, ed. James Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. New York: Cambridge University Press pp. 115–40.

Gerber, Alan, and Neil Malhotra. 2008. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3(3):313–326.

Gerber, Elisabeth R., and Daniel J. Hopkins. 2011. "When Mayors Matter: Estimating the Impact of Mayoral Partisanship on City Policy." *American Journal of Political Science* 55(2):326–339.

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

Glass, Gene V. 1976. "Primary, Secondary, and Meta-Analysis of Research." *Educational Researcher* 5(10):3–8.

Gross, Justin H. 2014. "Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance." Forthcoming in *American Journal of Political Science*.

Hagen, Richard L. 1997. "In Praise of the Null Hypothesis Significance Test." *American Psychologist* 52(1):15–24.

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

Hetherington, Marc, and Elizabeth Suhay. 2011. "Authoritarianism, Threat, and Americans' Support for the War on Terror." *American Journal of Political Science* 55(3):546–560.

Hill, Jr., Daniel W., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(3):1–27.

Hultman, Lisa, Jacob Kathman, and Megan Shannon. 2013. "United Nations Peacekeeping and Civilian Protection in Civil War." *American Journal of Political Science* 57(4):875–891.

Imai, Kosuke, Gary King, and Olivia Lau. 2008. "Toward a Common Framework for Statistical Analysis and Development." *Journal of Computational and Graphical Statistics* 17(4):892–913.

Jerit, Jennifer, Jason Barabas, and Scott Clifford. 2013. "Comparing Contemporaneous Laboratory and Field Experiments on Media Effects." *Public Opinion Quarterly* 77(1):256–282.

Kam, Cindy D., and Carl L. Palmer. 2008. "Reconsidering the Effects of Education on Political Participation." *Journal of Politics* 70(2):612–631.

Kam, Cindy D., and Elizabeth J. Zechmeister. 2013. "Name Recognition and Candidate Support." *American Journal of Political Science* 57(4):971–986.

Kinder, Donald R. 2007. "Curmudgeonly Advice." *Journal of Communication* 57(1):152–67.

King, Gary, and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9(2):137–163.

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

Kirk, Roger E. 1996. "Practice Signifance: A Concept Whose Time Has Come." *Educational and Psychological Measurement* 56(5):746–759.

Levendusky, Matthew S., and Michael C. Horowitz. 2012. "When Backing Down is the Right Decision: Partisanship, New Information, and Audience Costs." *Journal of Politics* 74(2):323.

Mutz, Diana C., and Byron Reeves. 2005. "The New Videomalaise: Effects of Televised Incivility on Political Trust." *American Journal of Political Science* 99(1):1–15.

Rainey, Carlisle. 2014. "Aruging for a Negligible Effect." Forthcoming in *American Journal of Political Science*.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11):1359–1366.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2014. "P-Curve: A Key to the File Drawer." *Journal of Experimental Psychology: General* 143(2):534–547.

Thompson, Bruce. 2001. "Significance, Effect Sizes, Stepwise Methods, and Other Issues: Strong Arguments to Move the Field." *Journal of Experimental Eduction* 70(1):80–93.

Thompson, Bruce. 2002. "What Future Quantitative Social Science Research Could Look Like: Confidence Intervals for Effect Sizes." *Educational Researcher* 31(3):24–31.

Tomz, Michael, Jason Wittenberg, and Gary King. 2003. "Clarify: Software for Interpreting and Presenting Statistical Results." *Journal of Statistical Software* 8(1).

Tomz, Michael R., and Jessica L. P. Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107(4):849–865.

Tukey, John W. 1962. "The Future of Data Analysis." *The Annals of Mathematical Statistics* 33(1):1–67.

Tukey, John W. 1991. "The Philosophy of Multiple Comparisons." *Statistical Science* 6(1):100–116.

Wainer, Howard. 1999. "One Cheer for Null Hypothesis Significance Testing." *Psychological Methods* 4(2):212–213.

Wonnacott, Thomas H., and Ronald J. Wonnacott. 1990. *Introductory Statistics*. 5th ed. New York: Wiley.

Yates, F. 1951. "The Influence of Statistical Methods for Research Workers on the Development of the Science of Statistics." *American Statistical Association Journal* 46(253):19–34.

Appendix

Meaningful Inferences

A Special Situations

In this paper, we lay out an approach that allows researchers to make strong empirical arguments that effects of interest are substantively significant. However, we would like to suggest several caveats.

Our goal is not to bring past research into question or to suggest that researchers should never test a simple directional hypothesis. Instead, our goal is to provide future researchers with a tool to strengthen their claims and more clearly communicate their evidence to readers in many situations, though not all. Indeed, while testing a null hypothesis of exactly no effect has received ample criticism from methodologists (Gill 1999, Gross 2014, and Hill and Jones 2014), some researchers defend its merits (e.g., Hagen 1997 and Wainer 1999). The method we propose, while powerful, has its own limitations.

In some situations, the scale or magnitude of the outcome or explanatory variable might not be interpretable, making the magnitude of effects difficult to discern. This can happen, for example, in observational studies in which the outcome or explanatory variables are measured poorly or in lab experiments, in which the experimental treatment might not map onto the real-world "treatment." In this situations, a simple sign-and-significance approach is a compelling alternative.

Measurement Error

In the special case of the normal-linear model, measurement error in the outcome variable in regression models simply increases the standard errors of the estimate and does not lead to

bias (i.e., the error term simply becomes $e_i + u_i$, where e_i is the usual residual and u_i is the measurement error for observation i). However, this specific finding does not generalize to other models. Consider a logit model for example. Suppose that certain events are randomly misclassified. This has the effect of shifting the probability of an event toward one-half because the outcome is more likely to be misclassified. As the probability of an event moves toward one-half, the marginal effects of all explanatory variables decrease. Thus, researchers must carefully consider the impact of measurement error in the outcome variable when making arguments about the magnitude of an effect.

Some researchers claim that measurement errors in explanatory variables attenuate estimated effects by biasing the estimates toward zero. This would work against researchers arguing for directional or substantively meaningful effects and lead to a conservative test. However, measurement errors in explanatory variables lead to attenuation only when the measurement error is random and uncorrelated with measurement error in the other explanatory variables. If measurement error exists in multiple variables and the errors are correlated, then the size and direction of the bias is quite difficult to discern. Researchers must carefully consider the impact of measurement error in the explanatory variables when making arguments about the magnitude of an effect.

What, then, should be done about measurement error? The first step is to carefully choose the best measures of the key theoretical concepts. The second step is to identify any potential measurement error. If measurement error is present and cannot easily be corrected, then the researcher should carefully discuss the biases that will likely results from these errors. In some cases, the bias might strengthen the claim, and in other cases, the bias might weaken the claim. The important point is to carefully consider the potential bias.

Lab Experiments

Lab experiments present a difficult environment for those arguing about the magnitude of an effect (as opposed to the direction). While some scholars, such as Gerber (2011), seem skeptical about the ability of lab experiments to determine the magnitude of an effect, Jerit, Barabas, and Clifford (2013) present evidence suggesting that lab experiments produces *larger* effects than similar field experiments. They suggest that the larger effects occur due to (1) forced exposure to the treatment (Gaines and Kuklinski 2011), (2) the pristine lab environment (Kinder 2007), (3) the obtrusiveness of the experiment (Webb et al. 2000), and (4) the time distance between the treatment and the outcome. These can combine to produce a smaller or larger effects in the lab, but Jerit, Barabas, and Clifford suggest that the effect should be larger on average.

In the case of lab experiments, the researcher should think carefully about how the study design maps onto the key real-world concepts. For example, is there a reason that a negative ad shown in the laboratory should have a similar size effect as an ad viewed at home after dinner? It seems plausible that the effect would be in the same *direction*, but the *magnitudes* might be quite different.

For example, Mutz and Reeves (2005), in studying the effects of incivility on political trust, use treatments that represent quite extreme versions of the level of civility (extreme politeness and calm) and incivility (disrespect, eye-rolling, raising voices) that we might find in actual campaigns. There are good statistical reasons to rely on treatments with large effects–it increases the power of the study–but this strategy prevents the researcher from drawing meaningful inferences about the magnitude of the effect. Similarly, an experimenter asking a subject to read a newspaper article might have very different effects than the simple publication of the identical article outside of the lab environment.

This does not imply that lab experiments are overused or less important that other types of experiments. Indeed, the control that makes estimating magnitude of an effect difficult might

make discovering the direction of an effect easier. The important point is to recognize that the "effect" in a lab experiment might correspond to the "effect" in the real world only in direction. Accordingly, the researcher should carefully consider this possibility and, if necessary, adjust the empirical claims accordingly, focusing on direction and not magnitude, in this situation, the usual directional hypothesis test maps onto the substantive claim perfectly.

B Replication of Kam and Zechmeister (2013)

Kam and Zechmeister (2013) argue that name recognition increases a candidate's support directly, by increasing the candidate's approval, and indirectly, by informing voters about the candidate's viability. The authors present three lab experiments to demonstrate the causal link between their concepts of interest, but they use a field experiment to boost the external validity of the laboratory results and establish their substantive importance.

Through a clever design exploiting routes that parents must use to drop their kids off at school, Kam and Zechmeister expose half of parents in a particular geographic region to four yard signs displaying fictitious candidate Ben Griffin's name. The other half of parents are not exposed to any name and serve as a control group. The authors then survey the parents and ask them to indicate their top three choices for city council seats by choosing among five actual candidates and two fictional candidates (Ben Griffin, whose name appeared on yard signs and Milt Jenkins, whose name did not appear on any signs and who serves as a placebo).

The authors summarize their results:

> Did recognition spurred on by political yard signs increase support for Ben Griffin in the treatment group? To determine if this is so, we examine the extent to which survey respondents selected Ben Griffin as one of their top three choices for council. As shown in [their] Table 3, in the control condition, only 13.9% of respondents placed Ben Griffin among their top three choices, but in the treatment condition,

> 23.9% of respondents placed Ben Griffin among their top three choices. This 10 percentage point difference is sizable given the modesty of the treatment. In light of the small sample size, it is statistically significant at generous levels ($p \approx 0.13$, one-tailed) (p. 983).

They continue:

> [W]e can examine the rates of selection of the two fictitious names, within each condition. Among the treated subjects, 23.9% of subjects placed Ben Griffin in the top three set, but only 13.0% placed Milt Jenkins in the top three set, a statistically significant difference at $p < 0.09$, one-tailed. Among the control subjects, 13.9% placed Ben Griffin in the top three set, and the identical percentage, 13.9%, placed Milt Jenkins in the top three set. The results from this field study lend generalizability to the claim established in our laboratory studies: name recognition increases candidate support in low-information elections (p. 983)

But can we be confident that this effect is indeed "sizable"? Are small effects plausible given the data? Figure 4 shows the estimated effects and 90% confidence intervals. Notice that while the estimated effects are of borderline statistical significance, the estimated effects of about 10 percentage points are quite large. However, much smaller effects are plausible as well. Indeed, the authors cannot even reject the tiniest of effects with these data. While we agree with the authors that that an effect of ten percentage points is indeed "sizable," their data are also consistent with small, negligible effects. As such, these data do not offer compelling evidence for a substantively meaningful effect.

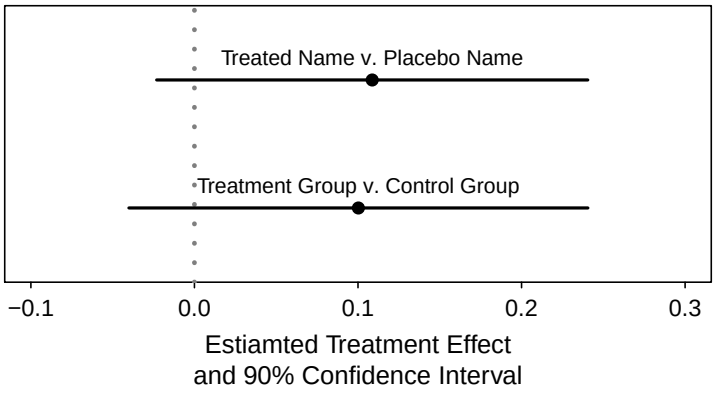

FIGURE 4: This figure provides the estimated treatment effect and 90% confidence intervals for placing candidate roadsigns along a street which citizens regularly drive on the probability of ranking the named candidate in the top three of seven candidates. The top estimate compares the treatment with the control group (i.e., parents driving along different routes) and the bottom estimate compares the named candidate to the placebo candidate among parents driving along the route with the yard signs.