Meaningful Inferences

The Importance of Explicit Statistical Arguments for Substantive Significance*

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" is only a small improvement over the much maligned "sign-and-significance" approach to interpreting statistical results. We argue that instead of interpreting the magnitude of a statistically significant effect, researchers should explicitly account for uncertainty when making judgments about substantive importance of statistical results. This requires the researcher to precisely define the effects that are and are not substantively meaningful and show that those effects are unlikely to generate the observed data. Using the effect of U.N. troops on civilian casualties during civil war and the effect of yard signs on candidate support, we show that our approach might validate or invalidate claims of substantive significance.

> 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, p. 13-14)

*We thank [many people]. The analyses presented here were conducted with R 3.1.0 and Stata 13. 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

[Introduction here.–Write later.]

Each of these questions require increasing levels of analysis and substantive interpretation and each receives decreasing levels of attention in political science research. As one might expect, political science research devotes less attention to the magnitude of the effects and their substantive importance. From 2011 to 2013, the *American Political Science Review* and the *American Journal of Political Science* published a total of 316 total articles, 73% of which were empirical analyses. Of this 73%, only about half of the articles discuss finding substantively meaningful. Furthermore, of the articles that discuss the substantive importance of their results, 83% simply assert that their results are meaningful and fail to make an explicit argument for their claim.

What We Want To Know

Most empirical research in political science focuses on estimating the effect Δ of an explanatory variable x on the expected value of an outcome of interest y. Formally, we might suppose that $E(y|x) = f(x)$ and define the "effect" or "quantity of interest" Δ as the difference between the average outcomes when x takes on a substantively meaningful high value and low value, so that $\Delta = E(y|x = x_{hi}) - E(y|x = x_{lo}) = f(x_{hi}) - f(x_{lo})$. Importantly, empirical work usually focuses on answering three fundamental questions about the effect Δ of x on y.

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

For example, consider the effects of U.N. troops on civilian casualties during a civil war. The first question concerns the direction of the effect. It could be that the troops have the intended effect and reduce casualties. However, the troops might also have the unintended consequences

of antagonizes the parties to the conflict and increasing civilian casualties. Once we establish the direction of the effect, we next want to know how large the effect is. Supposing a benefit, on average, does a commitment of 5,000 troops save 5, 50, or 500 civilians? Once we have established the magnitude of an effect, we want to make a substantive judgement about the importance of that effect. If committing the troops saved only 5 lives, then that might be a substantively unimportant effect since agencies would consider alternative strategies. However, if committing the troops save 500 lives, we would probably consider that a meaningful benefit and argue that the presence of U.N. troops leads to a substantively meaningful decrease in civilian casualties.

Establish the Sign

The first question that question that empirical research usually attempts to answer is the direction of the effect. Is the effect positive or negative?[1]

Suppose, for clarity, that the researcher offers a directional research hypothesis, suggesting that an effect of interest Δ is positive. Formally, we might denote this hypothesis as $H_r : \Delta > 0$. Then the researcher compares 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. Formally, we might write this as $H_0 : \Delta \leq 0$. To assess the evidence against the null hypothesis, the researcher usually calculates 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.[2] If this p-value is sufficiently small (by convention, less than 0.05), then the researcher rejects the null hypothesis in favor of the research hypothesis. In our example, the researcher would reject negative effects and conclude that the effect is positive. However, if the p-value is not sufficiently small, then the researcher declares that the data do

[1] 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.

[2] This is always the case in our sample.

not offer compelling evidence against the null and notes that the direction of the effect remains uncertain.[3]

Wilson and Piazza (2013) provide an example of this “sign-and-significance" approach. They hypothesize that both democratic and military regimes are more likely to experience terrorism than single-party regimes. They use a reference regime to compare for each model and discuss the implications in terms of the sign and significance of the coefficient. In particular, they find that “party-based autocracies are significantly more likely to have no terrorism in a given year," and that “this relationship is confirmed true below a 0.01 probability of error," (pp. 949). They also find, due to a p-value greater than 0.1, that the number of expected terrorist attacks in military regimes as compared to democracies are not statistically distinguishable. Though the authors do not, this may be an instance where it is possible to interpret this lack of statistical significance as ambiguous evidence to which we can make no claim rather than group the two regimes together.

There are three aspects of the sign-and-significance approach make note of. First, this style of argumentation is ubiquitous in political science. It is extremely rare to find empirical research in political science that does not perform a hypothesis test of some sort. Each empirical study in our sample performed some type of hypothesis test. Second, note that this approach is compelling not because it argues that the observed data are consistent with the researcher’s claim, but because the data are inconsistent with other claims. Third, notice that this argument for the direction of the effect explicitly takes into account the uncertainty of the estimated effect. If the uncertainty is large relative to the magnitude of the estimate then the researcher cannot (and usually does not) make confident claims about the direction of the effect of interest. However, if the uncertainty is small relative to the size of the estimated effect, then the researcher can draw confident conclusions about the direction of the effect.

[3]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).

Establish the Magnitude

Yet recent methodological work emphasizes that empirical work should go beyond estimating the direction of the effect (King, Tomz, Wittenberg 2000; Hanmer and Kalkan 2013; Gross 2014). In addition to the direction of an effect, the size of the effect matters as well.

However, interpreting magnitude is not always straightforward. Only some statistical models have naturally interpretable parameters. For example, a simple difference-in-means or normal-linear model have 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 these atypical situations, however, the researcher must do additional work to estimate a substantively meaningful quantity of interest.

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 not 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."

Yates continues:

> Tests of significance are preliminary or ancillary. The emphasis on tests of significance, and the consideration of the results of each experiment in isolation, have had the unfortunate consequence that scientific works have regarded the execution of a test of significance on an experiment as the ultimate objective. Results are significant or not significant and that is the end of it (p. 33).

Fortunately, recent conceptual work (King, Tomz, and Wittenberg 2000; Berry, DeMeritt, and Esarey 2010; and Hanmer and Kalkan 2013) and software development (Tomz, Wittenberg, King 2003; Imai, King, Lau 2007) empowers political scientists to move beyond a simple "sign-and-significance" approach and also present substantively meaningful measures of effect magnitude.

Establish the Substantive Importance

But ultimately, researchers 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."

For example, Kinne (2013) presents effects that he deems as substantively large enough to be important. He hypothesizes that there are three discernible ways that states network and form bilateral agreements and determines whether or not these are both statistically and substantively significant through his empirical analysis. Kinne interprets the substantive meaning of his results as states being "73% more likely to propose an agreement" or cooperate when there is a third party involved than any other combination of states and bilateral agreements (pp.

778). By determining which effects are substantively meaningful, he is able to make the claim that his research makes a substantial contribution to the literature on international cooperation. His analysis “puts endogenous influences in the spotlight, treating them not as methodological nuisances but as phenomena of substantive importance," and finds that typical much-cited cooperation influences “diminish in substantive significance once network influences enter the equation," (pp. 781).

Two Flawed Practices in Reasoning About Effect Magnitude

Two current practices dominate the way that researchers make statistical arguments for substantive significance. First, researchers might deem any statistically significant result as an important one. We refer to this as the “sign-and-significance” approach in which researchers simply note the sign of an effect and claim it is statistically significant. This approach is much maligned (e.g., Cohen 1990 and Gill 1999) and interprets the p-value as a measure of magnitude or importance.

The second practice, which is more common in political science, is to interpret the magnitude without explicitly accounting for the uncertainty of the estimate. We refer to this as the “magnitude-and-significance” approach in which researchers first check that an effect is statistically significant and then interprets the magnitude of the estimate. While the magnitude-and-significance approach is no doubt and improvement on the sign-and-significance approach, it retains many of its flaws

The Error of Sign-and Significance: Small p-Values Do Not Indicate Large Effects

A p-value is simply the probability of observing data at least as extreme as the observed data if the null hypothesis were true (e.g., Casella and Berger 2002 and DeGroot and Schervish

2002). All else equal, the p-value gets smaller as the effect size under investigation increases, but the p-value is also a function of the sample size (Gill 1999). 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. However, a minuscule p-value does not imply a large or important effect. "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 effect important, somewhat important, slightly important, or not at all important. The p-value is indirectly related to substantive significance at best. This disconnect occurs for two reasons. First, the p-value is only partly determined by the effect size–sample size plays a large part as well. Second, the absolute magnitude of the effect cannot be deemed substantively large, small, or negligible without the judgment of a substantive expert.

Current Practice Does Not Account for Uncertainty

The standard scientific practice in political science computes substantively interpretable estimates of effects and (1) determines whether these estimates are statistically significant and, if so, (2) makes a judgement about whether the magnitude is substantively important. [We need to add a cite here, but perhaps several examples will do. I'm confidence that this is the way it is done in practice, but we need to either point to a source that suggests the practice or illustrate with several examples that the practice is quite common.] This two step procedure has three unfortunate consequences. This tempts researchers to...

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

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

Consider, for example the hypothetical studies presented in Figure 1. If we considered 0.25 the minimal substantively meaningful effect, then only one study, Study A, seems to offer evidence for a substantively meaningful effect. Yet the current practice would also suggest that Study B offers evidence for a substantively meaningful effect as well. And while a large, meaningful effect is certainly plausible based on Study B, it fails to rule out small, negligible effects. In fact, the amount of evidence offers for the hypothesis of a meaningful effect by Study B is similar to that offered by Study C–neither study rules our small, negligible effects. Yet these studies are treated quite differently by the current practice in political science. In fact, the only improvement of the current practice over the sign-and-significance method is that the current practice managed to distinguish between Study D and Study A.

The key takeaway point is that it is important to apply similar standards of evidence to arguments for positive (or negative) effects and arguments for meaningfully positive (or meaningfully negative) effects. The usual logic of hypothesis testing requires that the researcher only declare that an effect is positive if and only if the evidence points overwhelming against negative effects. Similarly, researchers should not declare a positive estimate substantively meaningful simply because it is inconsistent with with negative effects and above some threshold. A better standard of evidence would require that researchers declare an effect meaningful if and only if it is inconsistent with negligible effects.

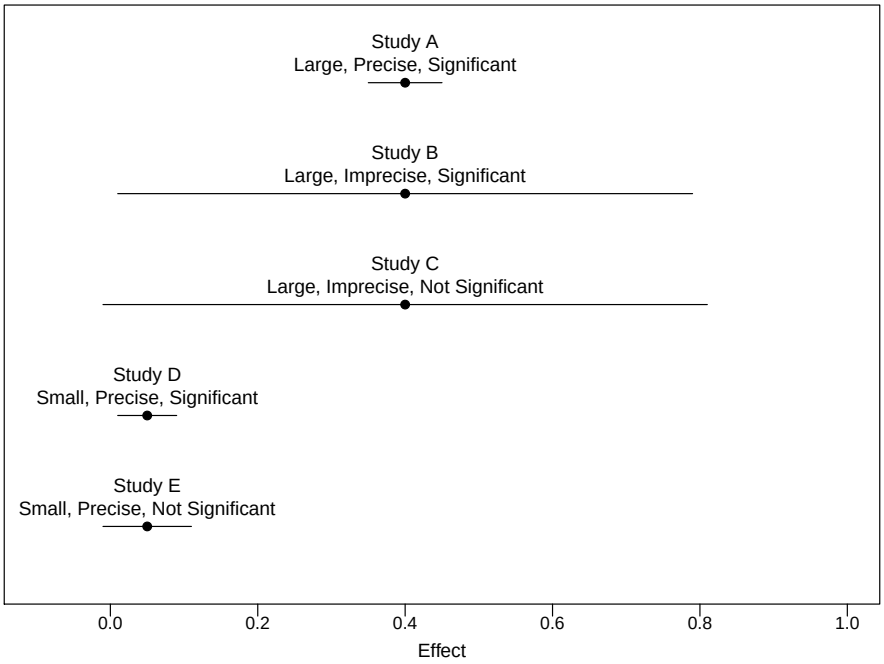

FIGURE 1: This figure provides several hypothetical studies to illustrate several points about arguments for substantive significance. Notice that Studies D and E produce similar results and Studies B and C produce similar estimates. If we take 0.25 as a the smallest substantively meaningful effect, then only Study A offers compelling evidence for a meaningful effect. Unfortunately, the current practice treats Studies A and B (and perhaps D) as similar evidence for a meaningful effect. See Table 1 for the typical interpretations.

A Compelling Argument for a Meaningful Effect

Researchers can make much more compelling and transparent arguments for meaningful effects by explicitly testing the claims we are making. For example, if a researcher claims that an effect is positive and substantively meaningful, then the research hypothesis is not $H_r : \Delta > 0$. Instead, it's $H_r : \Delta > m$, where m represents the researchers judgment about the smallest substantively interesting effect (or the largest substantively negligible effect). There is no need to jettison the entire hypothesis testing framework. In fact, the problems identified above can be addressed and the usual hypothesis testing framework can be kept intact if researchers are simply willing to apply their substantive judgment about which effects are and are not

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 provides the typical interpretations of the results in Figure 1 using the two common methods. Notice that the sign-and-significance methods and the current practice only manage to distinguish the evidence from Studies D and E. The current practice treats Studies A and B similarly and Studies C and E similarly. However, these studies offer vastly different amounts of evidence for or against an hypothesis of a meaningful effect.

meaningful to their hypotheses rather than the estimated magnitude.

The current practice in political science for arguing for a meaningful effect proceeds as follows:

1. Hypothesize that an effect is positive (or negative).
2. Test the hypothesis by computing a p-value.
3. If the null hypothesis can be rejected (i.e., $p < 0.05$), then claim support for the hypothesis and compute a substantively interpretable measure of the estimated effect size and make a substantive judgment about whether the estimate is substantively meaningful or not. If the null hypothesis cannot be rejected (i.e., $p < 0.05$), then note that the evidence is unclear and do not interpret the magnitude of the estimate.

The alternative approach that we propose proceeds as follows:

1. Hypothesize that the effect is positive and substantively meaningful (or negative and substantively meaningful).
2. Make a substantive judgment about the smallest substantively meaningful effect. That is,

choose an m such that m represents the smallest substantively meaningful effect.

3. Test the research hypothesis $H_r : \Delta > m$ using a conventional hypothesis testing approach.

Stating a Research Hypothesis of a Meaningful Effect

The first step is easy, but important. In order to make compelling arguments for meaningful effects, researcher must clearly state their claims. Rather than simply hypothesizing that an effect is positive or negative, researcher must make their claim that a variable has a meaningful effect from the outset. For example, Levendusky and Horowitz (2012) hypothesize that "bipartisan support in Congress for the president's policy should decrease audience costs." Notice that this hypothesizes a direction, but not a magnitude of the effect. Yet in discussing this hypothesis, Levendusky and Horowitz write:

> This type of unexpected (disconfirming) cue has an especially large effect on voters' decision-making processes (Baum and Groeling 2009; Eagly, Wood, and Chaiken 1978). In effect, it sends voters a strong signal that this is not a partisan decision, but rather a decision about what is best for the nation. Further, the fact that even the president's rivals supported his decision suggests to voters that the president did make the right call, which should lead all voters (regardless of partisan affiliation) to punish the president less harshly, thereby minimizing audience costs. [italics ours]

Thus, Levendusky and Horowitz carefully theorize about the magnitude of the effect, but only include the direction implied by the theory, and not the magnitude, into their hypothesis. They could improve the test of their theory by building the implication about the size of the effect directly into their hypothesis, predicting that "bipartisan support in Congress for the president's policy should substantially decrease audience costs." I have simply added "substantially" to their hypothesis. This provides a strong indication to readers (and the researcher) that the

theory implies the effect should be large and the researcher will provide a strong argument for a large negative effect, as opposed to simply a negative effect.

Choosing m

In order to make a compelling argument that an effect is substantively meaningful, the researcher must carefully define exactly which effects are and are not substantively meaningful. This quantity is the identical concept proposed by Rainey (2014), and as Rainey notes, political scientists should not insist on hard and fast rules for judging the effects that are and are not substantively meaningful.[4] Instead, we must insist that substantive scholars making substantive claims about politics also make substantive judgments about the importance of their effects. Thompson (2001) 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) writes: "[T]he 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 an author might reasonably disagree with the author about whether the effect size is noteworthy and then simply ignore the study."

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 are not at all transparent. Researchers do not fit one model and report the single p-value. Instead they fit many models and report

[4] 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.

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; Simonsohn, Nelson, and Simmons 2013; and Esarey and Wu 2014; see also Gelman and Loken 2014). Substantive judgments about effect sizes, however, require a large initial investment of careful thought to argue that certain effects are or are not substantively important. This judgment demands subjectivity. However, this subjective judgment is quite 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. Substantive scholars making substantive points about politics should be allowed and encouraged to make substantive judgments about magnitude (Achen 1982). Indeed, Kirk (1996) writes:

> [R]esearchers 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, hut in the name of objectivity, they are not expected or even encouraged to decide whether data are practically significant.(755).

Testing a Hypothesis of a Meaningful Effect

Once the researcher has clearly identified that set of effects that he considers substantively meaningful, the testing problem is straightforward. For an effect of interest Δ, the a researcher positing a "substantively meaningful, positive effect" must simply test her research hypothesis $H_r : \Delta > m$ against the null hypothesis $H_0 : \Delta \leq m$. For a researcher positing a "substantively meaningful, negative effect" must simply test her research hypothesis $H_r : \Delta < -m$ against the

null hypothesis $H_0 : \Delta \geq -m$.[5]

The case of t-statistics illustrates the parallel between testing for a meaningful positive effect and simply testing for a positive effect. If a researcher simply wishes to argue that an effect is positive (though perhaps substantively irrelevant), the t-statistic is given by $t = \frac{\Delta}{\sqrt{\widehat{Var}(\Delta)}}$. If a researcher wishes to argue that an effect is positive and substantively meaningful, then the required t-statistic is given by $t = \frac{\Delta - m}{\sqrt{\widehat{Var}(\Delta)}}$. The researcher can then use this t-statistic to compute p-values and determine if the respective null hypothesis of "no effect" or "a negligible effect" can be rejected.

Confidence Intervals

While the hypothesis testing framework is sometimes clear and convenient, confidence intervals offer even more information and are easier for readers (and researchers) to interpret. Specifically, the researcher simply needs to check that a 90% confidence interval contains values that are only consistent with the research hypothesis of a meaningful effect. Therefore, if the 90% confidence interval contains only large, meaningful effects, then the researcher can confidently reject small, negligible effects. 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.

A $100(1-\alpha)$% 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

[5] Though rarely used in practice, this the way hypothesis testing is introduced in in introductory textbooks (e.g., Wonnacott and Wonnacott 1990).

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 α and 90% and 95% confidence intervals respectively (see esp. Casella and Berger 2002, pp. 419-423).

Concretely, if the researcher predicts that an effect is positive and finds that the 90% confidence interval contains only positive effects, this is equivalent to rejecting the null hypothesis that the effect is less than or equal to zero at the 0.05 level.[6]

Replication of Hultman, Kathman, and Shannon (2013)

To illustrate how our approach can be used by a researcher to explicitly test their substantive claims, we replicated Hultman, Kathman, and Shannon's (2013) assessment of civilian protection by UN peacekeeping operations (PKOs). Hultman, Kathman, and Shannon offer a theory that civilians can be successfully protected by UN PKOs when those missions are composed of military troops and police in adequately large numbers. Following convention, these authors present their empirical results by (1) showing that the relevant quantity of interest is statistically significant and (2) arguing that the estimated effect is substantively meaningful.

They argue that PKOs mitigate violence both on the battlefield and behind the battlefield's frontlines for a variety of reasons while the UN's ability to intervene is contingent upon the size and personnel composition of the deployment. To evaluate this argument, Hultman, Kathman, and Shannon hypothesize that as the UN commits both more military troops and more police personnel to a conflict, the amount of violence committed against civilians will decrease. Because their original hypothesis is directional, we can use two one-tailed tests, as detailed in the previous section, to check whether or not the 90% confidence interval contains values that are consistent only with a hypothesis of a meaningful effect.

[6] Similarly, if the researcher predicts that an effect is *negative* and finds that the 90% confidence interval contains only *negative* effects, this is equivalent to rejecting the null hypothesis that the effect is *greater* than or equal to zero at the 0.05 level (see Freedman, Pisani, and Purves 2007 pp. 383-385 for more on this point and DeGroot and Schervish 2012 pp. 485-493 for an alternative perspective).

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)

However, 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 these plots in Figure ??.

They write:

> 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)

They argue that this reduction from an average of about 106 to about 2 is substantively important.

> 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).

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

Following their use of a negative binomial model, we replicate their results, calculate first-differences, and 90% confidence intervals around the change in the expected civilian deaths as UN military troops increases. Figure 2 shows these confidence intervals. At an expense of $2 million, or roughly 2,000 troops, leads to about 65 fewer civilian casualties, on average. However, the data strongly 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, effects small than about 70 fewer civilian deaths are inconsistent with these data and the effect might be as large as 140 fewer deaths.

In this case, the authors have strong evidence for a meaningful effect, even after uncertainty is taken into account. The authors are able to confidently reject trivial effect sizes because the confidence intervals contain only meaningful effects.

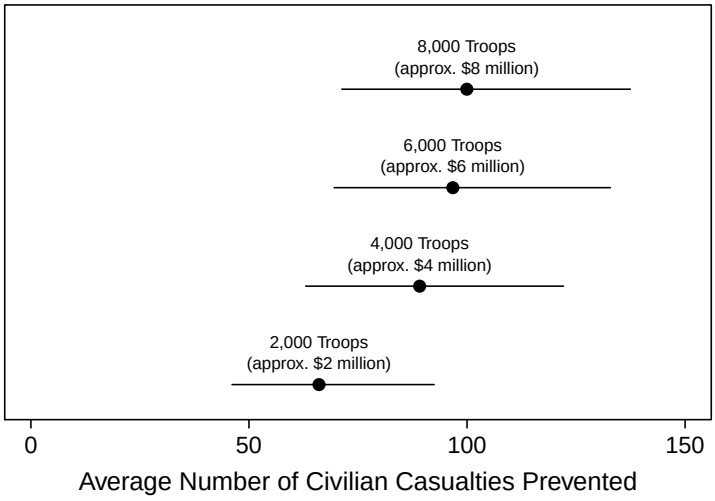

FIGURE 2: This figure provides first differences, in dollar and troop amounts, and their corresponding number of prevented civilian casualties.

Replication of Kam and Zechmeister (2013)

Kam and Zechmeister (2013) argue that candidate name recognition by directly, by increasing candidates approval, and indirectly, by informing voters about the candidate's viability. The authors present three lab experiments to demonstrate the cause link between their concepts of interest, but they use a field experiment to boost the external validity of the laboratory results. Through a clever design exploiting routes that parent 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 a fictitious candidate Ben Griffin's name. The other half of parents are not exposed and serve as a control group. The authors then surveyed the parents and asked them to indicate their top three choices for city council seats by choosing among five actual candidates and two fictional candidates (one whose name appeared on yard signs and a second, placebo, whose name did not appear on any signs). 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).

But can we be confident that this effect is indeed "sizable"? Are small effects plausible given the data? Figure 3 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 the estimated effects are "large and significant," these data do not offer compelling evidence for a substantively meaningful effect.

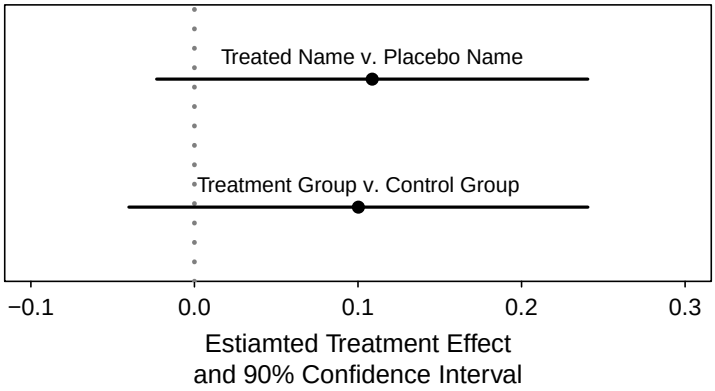

FIGURE 3: This figure provides the estimated treatment effect and 90% confidence intervals for placing candidate roadsigns along a street which citizens regularly drive on the probability of ranking the named candidate in the top three of seven candidates. The top estimate compares the treatment with 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.

Discussion

In this paper, we have laid 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 *some* situations. Indeed, while testing a null hypothesis of exactly no effect has received ample criticism from methodologists (Gill 1999, Gross 2014, Hill and Jones 2014), some researchers

defend its merits (e.g., Hagan 1997 and Wainer 1999 I might want to add Abelson1997 here as well). 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.

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 model. 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 more outcome is more likely to be misclassified. Of course the logit link function is non-linear and as the probability of an event moves toward 0.5, the marginal effects of all explanatory variables increase. Thus, researchers must carefully consider the impact of measurement error 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. The argument goes that measurement error works against researchers arguing for directional or substantively meaningful effects. For example, [add an example here]. 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.

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 present a difficult environment for those arguing about the magnitude of an effect (as opposed to the direction). While some scholars seem skeptical about the ability of lab experiments to determine the magnitude of an effect (Gerber 2011), Jerit, Barabas, and Clifford (2013) present evidence suggesting that lab experiments produces *larger* effects than similar field experiments. They suggests 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 careful about how the study design maps onto the key real-world concepts. For example, is their reason that a negative add shown in the laboratory has an effect similar in magnitude as an add viewed at home after dinner? It seems plausible that the effect occurs 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 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, 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.

Conclusion

In this paper, we have argued that researchers should explicitly test their substantive claims. In some cases this requires testing a directional hypothesis. In other cases, this requires an explicit test for a meaningful effect. The suggestion for empirical arguments that effects lie above a threshold of substantive significance is not new (Achen 1982, Gross 2014; see also Esarey and Danneman 2014 and Rainey 2014). However, explicit testing of substantive claims is not yet common practice and scholars rarely offer complete, substantive interpretations of the range of effects contain in confidence intervals. The current practice continues to be testing a directional research hypothesis and interpreting the substantive significant 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. We hope that researchers being to make precise claims about substantive significance and offer compelling evidence for those claims using confidence intervals of explicit tests for meaningful effects. Using this approach, researchers will (1) compute quantities that are of direct substantive interest, (2) clarify there 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. The result will be more transparent

substantive claims and clearer communication of the empirical evidence for these claims.