Political Science Research and Methods Vol 3, No. 1, 95–111 January 2015
© The European Political Science Association, 2014 doi:10.1017/psrm.2014.14

A Quantitative Method for Substantive Robustness Assessment

JUSTIN ESAREY AND NATHAN DANNEMAN

Empirical political science is not simply about reporting evidence; it is also about coming to conclusions on the basis of that evidence and acting on those conclusions. But whether a result is substantively significant—strong and certain enough to justify acting upon the belief that the null hypothesis is false—is difficult to objectively pin down, in part because different researchers have different standards for interpreting evidence. Instead, this article advocates judging results according to their "substantive robustness," the degree to which a community with heterogeneous standards for interpreting evidence would agree that the result is substantively significant. This study illustrates how this can be done using Bayesian statistical decision techniques. Judging results in this way yields a tangible benefit: false positives are reduced without decreasing the power of the test, thus decreasing the error rate in published results.

A long and cross-disciplinary literature stresses the importance of assessing the substantive significance of empirical results (Achen 1982; Hunter 1997; McCloskey 1998; Gill 1999; Lunt 2004; Ziliak and McCloskey 2004, 2008; Miller 2008; Siegfried 2010). We say a result is "substantively significant" when it is justifies acting on the belief that the null hypothesis is false.[1] Naturally, substantive significance of this sort is partially a matter of individual judgment. These standards include how averse a researcher is to drawing a mistaken conclusion and how large a relationship must be in order to be scientifically or politically meaningful. Thus attempts to use decision theory (Wald 1950; DeGroot 2004 {1970}; Pratt, Raiffa and Schlaifer 1996; Manski 2007) to establish criteria for substantive significance are likely to degenerate into disagreement over the utility function chosen to encapsulate such standards.

In this article, we argue that it is more helpful to assess a result's substantive robustness—the degree to which a community with heterogeneous standards for interpreting evidence would agree that the result is substantively significant—rather than whether it meets any individual standard. This focuses attention away from contention about which specific standards are appropriate and onto the breadth of evaluation standards that can be satisfied by a particular piece of evidence. The idea is to enable a researcher to objectively demonstrate whether his/her

Justin Esarey is Assistant Professor of Political Science, Rice University, PO Box 1892, Houston, TX 77251-1892, USA (justin@justinesarey.com). Nathan Danneman is Data Scientist, L-3 Data Tactics, 7901 Jones Branch Dr. No. 700, McLean, VA, USA (ndanneman@gmail.com). This research was supported in part by a University Research Committee grant from Emory University. We thank Bill Berry, Jeff Staton, Tom Clark, Drew Linzer, Emily Ritter, John Gasper, Stu Jordan, Tim Salmon, David Siegel, Jackie DeMeritt, Deirdre McCloskey, Andrew Martin and respondents at presentations of this research at Emory University, Florida State University, Georgia State University, the University of Rochester, the 2010 Annual Meeting of the Society for Political Methodology, the 2010 Annual Meeting of the Midwest Political Science Association, and the 2009 Annual Meeting of the American Political Science Association for helpful comments and criticism. To view supplementary material for this article, please visit http://dx.doi.org/10.1017/psrm.2014.14.

[1] We thank an anonymous reviewer for suggesting this phraseology.

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

results should be regarded as substantively significant by a scientific community, even if there is significant disagreement within that community over what a substantively significant result looks like. This approach is compatible with any number of sufficiently flexible utility functions, but we offer one that we believe is well suited to the task.

We also offer evidence that requiring results to be substantively robust as well as statistically significant could improve the quality of hypothesis tests in the discipline. Our Monte Carlo studies reveal that:

1. Results that are both statistically significant and substantively robust are less likely to be false positives[2] than those that are merely statistically significant, with no loss in power.[3]
2. Statistically significant but substantively meaningless results are extremely unlikely when a scientifically important relationship exists; observing this pattern allows us to confidently conclude that the result is a false positive.

In short, substantive robustness tests may enable researchers to discriminate between genuine results and statistical anomalies.

As has been noted in the past (Ziliak and McCloskey 2004), the statistical significance of a result can be markedly more *or* less robust than its substantive significance. We illustrate this by re-analyzing two recently published results. For Clinton (2006), we find that some effects that are statistically significant are not substantively robust to a reasonable range of standards for the interpretation of evidence. We also re-analyze data from Clark and Golder (2006), confirming the substantive robustness of the authors' results while demonstrating the application of our technique in situations in which multiple variables determine the effect of interest (such as in models with interacted variables).

DRAWING CONCLUSIONS FROM EVIDENCE

We first present an example to help the reader crystallize the dimensions of judgment on which researchers may differ when drawing conclusions about substantive significance. We think this is easiest to do using an analogy between an intuitively compelling (but evidence-based) decision and a more ordinary social scientific decision that a researcher might make in day-to-day work.

Figure 1 illustrates this analogy. Let us assume that researchers develop an interesting new drug that can reliably produce any level of weight loss desired by the user. But researchers suspect that a fatal cancer *might* be a side effect of the drug. Figure 1a depicts a hypothetical posterior belief distribution of the change in cancer probability based on a series of (imagined) studies; we denote this change in cancer probability r. The distribution is centered over a 25 percent increase in a subject's chance of developing cancer, but the 95 percent confidence interval is [−48 percent; 98 percent] corresponding to a one-tailed p-value of about 0.25. That is, the relationship between this drug and cancer is statistically insignificant.

Would you take this drug?

A rational chooser[4] would consider the potential consequences of taking the drug vs. not taking the drug under each of the possible states of the world—each of the possible values of

[2] By "false positive," we mean incorrectly rejecting a one-sided null hypothesis when that hypothesis is true.

[3] By "power," we mean the probability of rejecting a one-sided null hypothesis when that hypothesis is false.

[4] See DeGroot (2004 {1970}, Chapter 7), French and Insua (2000, Chapter 6), and Pratt, Raiffa and Schlaifer (1996, Chapters 3 and 4) for general overviews of Bayesian statistical decision theory.

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

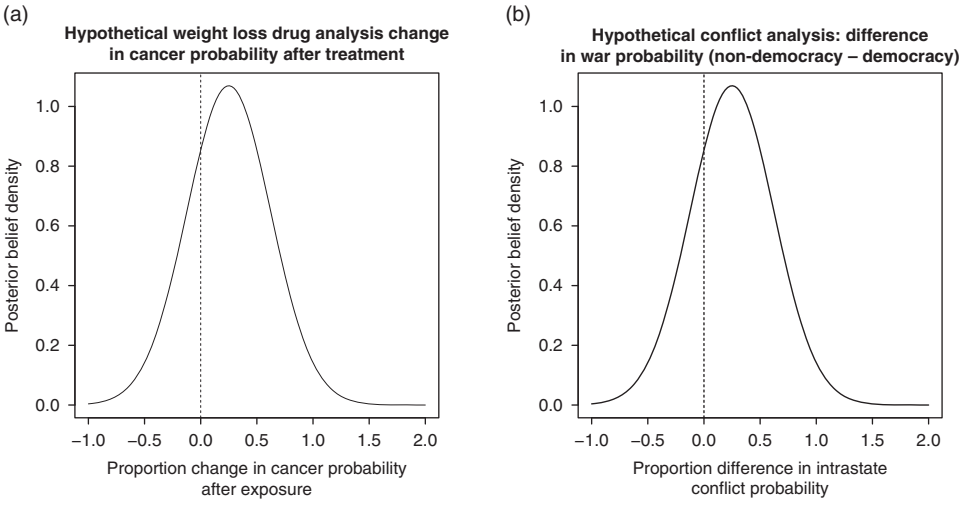


Fig. 1. Hypothetical evidence-based decisions (a) Would you take this weight loss drug? (b) Would you believe that non-democracies have more civil wars than democracies?

$\partial \Pr(\text{cancer})/\partial \text{take drug} = r$ under the posterior—and then determine whether taking the drug has a positive expected utility:

$$E[u(\text{take drug}) - u(\text{don't take drug}) \mid \text{data}] =$$

$$\int [u(\text{take drug} \mid r) - u(\text{don't take drug} \mid r)] f(r \mid \text{data}) dr =$$

$$\int u(\text{take drug} \mid r) f(r \mid \text{data}) dr - \int u(\text{don't take drug} \mid r) f(r \mid \text{data}) dr =$$

$$E[u(\text{take drug}) \mid \text{data}] - E[u(\text{don't take drug}) \mid \text{data}]. \quad (1)$$

The decision to choose to take the drug when the expected utility (in Equation 3) is greater than zero, and to not take it otherwise, is the *Bayes decision rule*:

$$d(x) = \text{take drug if } E[u(\text{take drug}) \mid x] > E[u(\text{don't take drug}) \mid x]$$

$$\text{don't take drug otherwise}$$

where we let x signify the data. The rule associates each state of the world $f(r \mid \text{data})$ with an optimal *Bayes action* that maximizes expected benefit (French and Insua 2000, 148). Uncertainty about the actual effect of the drug plays a role in the decision[5] through $f(r \mid \text{data})$.

[5] The "Bayesian" in Bayesian statistical decision theory comes from the fact that these conditions rely on the posterior probability of r given the data, $f(r \mid \text{data})$, which is determined by Bayes' rule:

$$f(r \mid \text{data}) = \frac{f(\text{data} \mid r) f(r)}{\int f(\text{data} \mid r) f(r) dr}.$$

The analyst must specify a prior, $f(r)$, before determining whether these conditions are met; we will discuss prior specification below.

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

This Bayes decision rule $d(x)$ is optimal in that it minimizes the *Bayes risk*, the expected loss in utility $l(\bullet) = -u(\bullet)$ from a decision rule averaging over both sampling variation in the data and the prior distribution of r:

$$BR(d) = \int \left[\int l(d(x) \mid r) f(x \mid r) dx \right] f(r) dr.$$

Summarily, it is rational to commit to following the Bayes decision rule because it maximizes utility (equivalent to minimizing loss) given all sources of uncertainty in our empirical evidence.[6]

While there may be disagreements over the quality of the evidence, we set them aside to focus on differences in interpretation of the evidence. These differences in judgment presumably come through the utility function u. For example, some people may be more loss averse[7] than others; they would presumably be less willing to gamble on the possibility of getting cancer in order to lose weight. The difference on this dimension is in how people weight the *relative* value of forsaking a gain if the drug is harmless compared to the value of avoiding a loss if it causes cancer. People may also differ in their *absolute* valuation of a cancer outcome: a small but completely certain increase in cancer incidence (say, 0.5 percent) might be negligible to some people compared to the benefit of weight loss, while others may consider this incidence extremely meaningful from a personal or public health perspective.

Now consider another decision: should we act on the conclusion that democratic states are less likely to suffer a civil war compared to non-democracies, or treat democracies as being at least equally likely to experience a civil war? Based on the (hypothetical) evidence shown in Figure 1b, the posterior distribution of the difference in intrastate conflict probability (which we denote as β) is the same as the distribution of the difference in cancer probability from Figure 1a. As before, a rational chooser should determine whether there is a greater scientific (or policy) benefit to acting on the conclusion that democracies are less susceptible than non-democracies to civil war, or to treating democracies as at least equally susceptible, based on the

[6] A proof of this proposition is given by French and Insua (2000, 164–5). Begin with the expression of Bayes risk:

$$BR(d) = \int \left[\int l(d(x) \mid r) f(x \mid r) dx \right] f(r) dr.$$

Noting that the central term $R(x, \beta) = \int l(d(x) \mid r) f(x \mid r) dx$ is often referred to as the (frequentist) *risk*, the expected loss integrating over sampling variation given a fixed state of the world r. Reverse the order of integration without loss of generality:

$$= \int \left[\int l(d(x) \mid r) f(x \mid r) f(r) dr \right] dx. \tag{2}$$

This implies that a decision rule that minimizes the expected loss for every particular dataset, $\int l(d(x) \mid r) f(x \mid r) f(r) dr$, will also minimize the Bayes risk. But this is precisely what the Bayes decision rule prescribes; it minimizes:

$$\int l(d(x) \mid r) f(r \mid x) dr = \frac{\int l(d(x) \mid r) f(x \mid r) f(r) dr}{\int f(x \mid r) f(r) dr},$$

where the numerator is the central term of Equation 2 and the denominator is a constant.

[7] Here, *loss aversion* refers to overvaluing reductions in utility relative to gains when computing expected values; this is not directly related to the similarly named *loss function* (i.e., the negative of the utility function in Equation 3).

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

available evidence:

$$E[u(\text{treat as unequal}) - u(\text{treat as equal}) \mid \text{data}] =$$

$$\int [u(\text{treat as unequal} \mid \beta) - u(\text{treat as equal} \mid \beta)] f(\beta \mid \text{data})] d\beta. \quad (3)$$

More generally, we could write this decision as:

$$E[u(\text{act on alternative}) - u(\text{act on null}) \mid \text{data}] =$$

$$\int [u(\text{act on alternative} \mid \beta) - u(\text{act on null} \mid \beta)] f(\beta \mid \text{data})] d\beta, \quad (4)$$

where the alternative hypothesis is that democracies are less prone to civil war, and the (default) null hypothesis is that they are at least equally prone to civil war. Of course, we would not expect $u(\text{act on alternative} \mid \beta) - u(\text{act on null} \mid \beta)$ to be the same for this decision as for the decision to take the weight loss drug. But the considerations involved in shaping the judgment are similar: (1) how much does the scientist weight the consequences of falsely acting on the alternative against the consequences of falsely acting on the null, and (2) how large must democracies' advantage in civil war susceptibility be before it merits scientific and political attention?

Rather than argue for a particular answer to these questions, our goal is to determine whether researchers with a variety of different standards would make the same decision according to Equation 4. We must still choose a utility function, but ideally it will be one that can fairly represent a variety of standards for evaluating evidence. In particular, it should be able to represent a continuum of relative valuations for false positives (mistakenly acting on the alternative hypothesis) and false negatives (mistakenly acting on the null hypothesis).

QUANTIFYING THE DECISION: UTILITY, LOSS AVERSION AND DECISION RULE

The functional form of $u(\text{act on alternative} \mid \beta) - u(\text{act on null} \mid \beta)$ is the encapsulation of scientific judgment in this framework, and there is no universally correct choice. Recall that our objective is to find whether different standards of judgment would lead to the same conclusion about the substantive significance of an empirical result. Thus the precise choice of functional form is less important than its ability to approximate a diversity of preferences about (1) the trade-off between false positives and false negatives, and (2) the minimum effect size required for substantive significance. We suggest a function that is simple, yet flexible enough to adapt to varying decision contexts and the preferences of different researchers. We do not claim that it is unique or superior to all possibilities, but it meets our stated criteria well and provides some inferential benefits that we explore below.

For the decision of whether to act on the conclusion that the relationship between two variables (such as democracy and intrastate conflict) is positive and large enough to be scientifically or politically important, we focus on a loss-averse utility function:

$$u(\text{act on alternative} \mid \beta, \gamma, c) - u(\text{act on null} \mid \beta, \gamma, c) = \gamma^{-\text{sign}(\beta - c)}[\beta - c]. \quad (5)$$

Here, the alternative is the conclusion that $\beta \geq c$ and the null is the conclusion that $\beta < c$.[8] This utility function is depicted in Figure 2 for $c = 0$. In terms of our previous democracy-conflict example, this is equivalent to saying that a non-democracy's chance of suffering a civil conflict is at least c more than a democracy's. Utility is gained in proportion to distance from the

[8] The utility of accepting an alternative hypothesis of a negative relationship uses the same function, but replaces β with $-\beta$ and c with $-c$ so that negative β values yield positive utility and the substantive threshold for significance is less than zero.

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

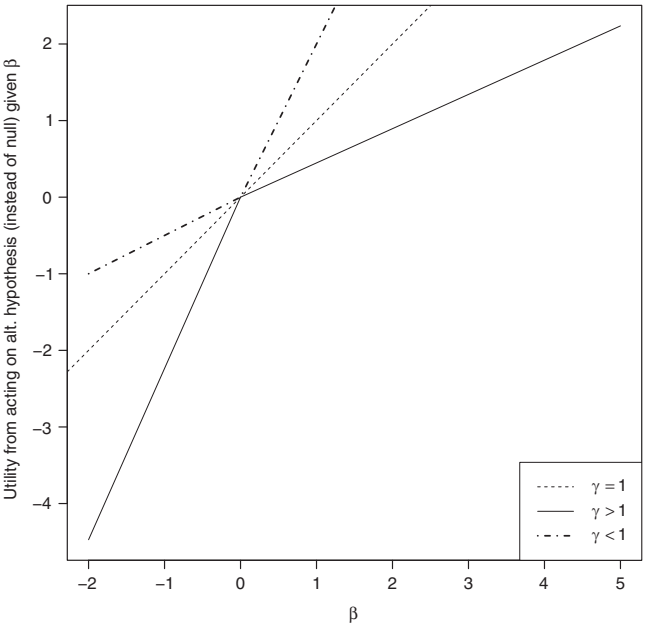


Fig. 2. Loss-averse utility functions

threshold for substantive significance c. With all else held equal (that is, in a single decision context), bigger relationships are more substantively important than small ones, and the smallest positive relationship worth paying attention to is of size c. That is, the consequences of accepting this conclusion are linearly related to the size of the difference. γ is a parameter that describes the valuation ratio for correct and incorrect decisions, or the degree of loss aversion that the researcher exhibits; it is the degree of kink in the function in Figure 2. Downward kinks indicate that correctly acting on the alternative is less important than incorrectly doing so (a false negative is less damaging than a false positive), and correspond to $\gamma > 1$. Upward kinks indicate that correctly acting on the alternative is *more* important than incorrectly doing so (that is, a false positive is less damaging than a false negative) and correspond to $\gamma \in (0, 1)$.

The Bayes decision rule under this utility framework is to act on the alternative hypothesis ($\beta \geq c$) whenever:

$$\int \gamma^{-\text{sign}(\beta - c)}[\beta - c] f(\beta \mid \text{data}) d\beta > 0 \tag{6}$$

and to act on the null hypothesis otherwise.[9] There are two unknowns, γ and c, which correspond to the two variable dimensions of scientific judgment that we laid out at the start.

[9] Letting x stand for the data and $\Delta(x) = \int \gamma^{-\text{sign}(\beta-c)}[\beta - c] f(\beta \mid x) d\beta$ from equation (6), the frequentist risk $R(x, \beta) = \int l(d(x) \mid \beta) f(x \mid \beta) dx$ of this rule is

$$R(d, \beta) = \int [\gamma^{-1}[\beta - c] * I(\beta > c) * \Pr(\Delta(x) < 0 \mid \beta) - \gamma[\beta - c] * I(\beta \leq c) * \Pr(\Delta(x) > 0 \mid \beta)] dx$$

Where we set $u(\text{null} \mid \beta < c) = u(\text{alternative} \mid \beta \geq c) = 0$, $u(\text{null} \mid \beta \geq c) = -\gamma^{-1}[\beta - c]$, and $u(\text{alternative} \mid \beta < c) = \gamma[\beta - c]$; recall that $l(\bullet) = -u(\bullet)$. The Bayes risk further integrates over the distribution of β using the decision rule that forms the basis for R: $BR(d) = \int R(x, \beta) f(\beta) d\beta$ Given that the decision rule is (by construction) a Bayes decision with respect to the prior distribution $f(\beta)$, it minimizes the Bayes risk and is also f-admissible (French and Insua 2000, Proposition 30 on p. 166; see also p. 144)

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

By determining the c^* that solves:

$$\int \gamma^{-\text{sign}(\beta - c^*)} [\beta - c^*] f(\beta \mid \text{data}) d\beta = 0, \tag{7}$$

we can sketch a γ/c^* curve over which the researcher is indifferent to acting on the alternative hypothesis given γ and c^*.

COMPUTING C^* AND CHOICE OF A PRIOR

The vast majority of mainstream statistical software packages can compute the γ/c^* curve using add-on packages. We have developed software for R and Stata that, when given a γ, will compute a c^* immediately after a linear regression or other generalized linear model. The software will also determine c^* when given manually entered information from a published table (without the accompanying dataset).[10]

The software package assumes a truncated uniform prior distribution $f(\beta)$ on the interval defined by $\hat{\beta} \pm 8\hat{\sigma}$. This structure assumes minimalistic knowledge of the underlying parameters before examining a dataset. It also allows for a Bayesian interpretation of frequentist results: for the classical linear regression model, $f(\beta \mid \text{data})$ takes a multivariate t distribution with $n-k$ degrees of freedom (Gelman *et al*. 2003, 355–7). The software therefore uses a t distribution for the posterior when computing c^* in these cases. For generalized linear models with this prior estimated via maximum likelihood, $f(\beta \mid \text{data})$ is asymptotically normal as $n \to \infty$ (Gelfand and Ghosh 2000, 4–8), and therefore the software uses the normal distribution in these instances.

Computing c^* is a matter of finding the roots of Equation 7 (or the equivalent for an alternative utility function). This class of problem is already solved via iterative maximization algorithms, such as Newton-Raphson,[11] but we are required to compute an integral at each iteration because we are calculating an expected value rather than a maximum or modal value. For results with a t-distributed or normally distributed posterior, we can use standard quadrature approaches to quickly compute the integral for each step of the maximization process.

Some analysts may wish to impose stronger prior beliefs, which will change the distribution of $f(\beta \mid \text{data})$. If the prior is conjugate with the posterior, the analyst must replace the multivariate t or normal distribution with the appropriate posterior distribution before calculating c^*. For more exotic posteriors, where $f(\beta \mid \text{data})$ is difficult to analytically express, an approximation can be computed via Markov Chain Monte Carlo methods, stored, and then used in the integration of Equation 7.

WHAT DOES A SUBSTANTIVELY ROBUST (OR NON-ROBUST) FINDING LOOK LIKE?

Figure 3b sketches the γ/c^* curve for our democracy and war example using Equation 7. The relevant question is: can we conclude that non-democracies are susceptible to a greater risk of civil war on the basis of the evidence of Figure 1b? Our answer is that a researcher would accept an increased civil war probability under the point on the curve corresponding to his/her γ. Equivalently, a researcher would decide how large a relationship had to be before it was scientifically or politically meaningful and how loss averse s/he is, then find the appropriate

[10] We have made this software package available at http://dx.doi.org/10.7910/DVN/25791.

[11] Many maximization algorithms involve finding a root of $\frac{df(x)}{dx}$ to find a maximum of $f(x)$. The root-finding capability of these algorithms is easily adapted to non-maximization root solutions.

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

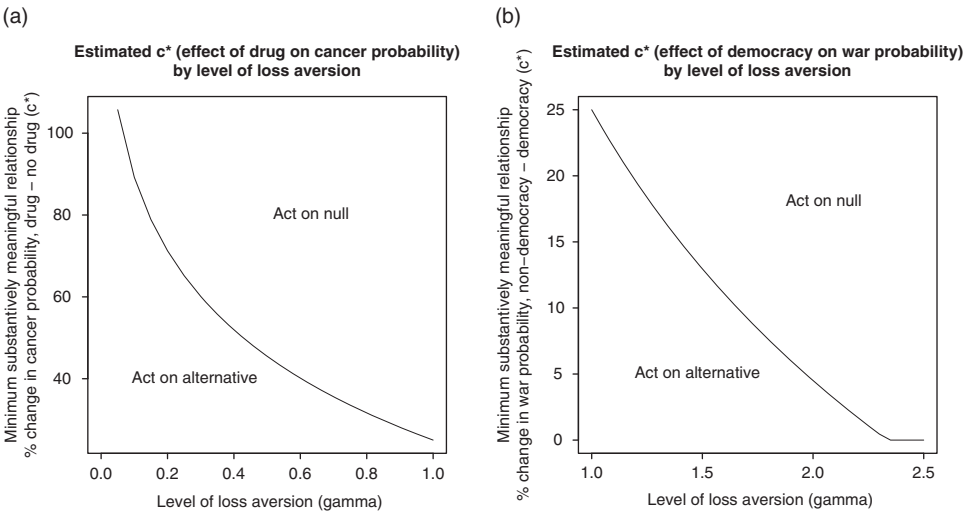


Fig. 3. Substantive robustness of inference from Figure 1 (a) Drugs and cancer probability (Figure 1a) (b) Democracy and war probability (Figure 1b)

point in Figure 3b and make the decision indicated. For example, a person who weighted false positives and false negatives equally (with $\gamma = 1$) would be willing to act on the belief that non-democracies are $\leq$25 percent less likely to experience civil war than comparable non-democracies. But a person who valued false positives four times as much as false negatives ($\gamma = \sqrt{4} = 2$) would only be willing to act on a $\leq$5 percent difference in civil war risk, and a person who valued false positives even more highly would always act on the null (of a nonexistent or negative relationship). In short, the association between democracy and civil war risk (in this hypothetical example) is not robust to reasonable ranges of variation in standards for interpreting evidence. It is not necessarily important for a researcher to know his or her own γ, but simply to acknowledge that a community of people with reasonable variation in how much they valued false positives more highly than false negatives would not agree on the substantive significance of the relationship between democracy and civil war risk.

By contrast, the conclusion that we should *not* take the weight loss drug (for fear of cancer risk, as shown in Figure 1a) is quite robust. Figure 3a sketches the γ/c^* curve for this example; the key difference, compared to the democracy/war example, is that $\gamma \in [0, 1]$ would be appropriate for a person who was much less concerned about mistakenly acting on the alternative (accidentally concluding that the drug causes cancer when it does not, and therefore *not* taking the drug when it would be beneficial) compared to mistakenly acting on the null (mistakenly believing that the drug does not cause cancer when it actually does, and therefore taking the drug when it increases one's likelihood of developing cancer). For this reason, even wide variation in standards for the interpretation of evidence produces the conclusion that there is a large and substantively meaningful link between the drug and cancer risk. It is not so important for any particular person to locate their personal value of γ, but simply to acknowledge that it lies somewhere between 0 and 1 (that is, that they value false negatives more highly than false positives in some measure) and that people with diverse preferences in this domain would avoid taking the drug.

We think these examples illustrate the value of focusing on the *robustness* of substantive significance to different standards rather than on any particular judgment of substantive significance.

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

Figure 3 makes clear that comparatively small differences in loss aversion (that is, aversion to falsely rejecting the null) can sometimes (but not always) translate into large differences in the assessment of substantive significance. Aggressive researchers who are relatively indifferent to false positives ($\gamma \approx 1$) would conclude (based on our hypothetical evidence) that there is a substantively significant association between democracy and civil war probability as long as they thought that a 25 percent increase in war probability was meaningful. More cautious researchers, who weight false positives as more damaging ($\gamma \gtrsim 2.5$), would *never* draw this conclusion. These are rather wide differences in judgment that relate to incommensurable differences in scientific viewpoint. Consequently, we think that arguing for a universally acceptable γ, let alone a universally acceptable form for $u(\text{act on alternative}) - u(\text{act on null})$, is probably hopeless in the context of academic research. By focusing on the breadth of preferences for which a result is substantively significant, rather than on any particular preference, we sidestep the knotty question of "whose judgment is correct?" in favor of the more readily answerable "would this evidence satisfy most researchers?" For the example of the weight loss drug, researchers with a wide variety of standards would all agree that the drug should not be taken.

Substantive Robustness Assessment Can Improve the Quality of Published Results

When it comes to statistical hypothesis testing, we determine that it is advantageous to require a result to be both substantively robust and statistically significant in order to be scientifically notable, compared to requiring statistical significance alone. Aside from the theoretical and descriptive value of assessing the substantive robustness of results, requiring a statistical relationship to be statistically significant *and* substantively robust before rejecting the null hypothesis (of no relationship) results in a more powerful test (a greater probability of rejecting false null hypotheses) at any given size (probability of rejecting correct null hypotheses). This advantage accrues to the combined procedure because results that are statistically significant but substantively fragile frequently occur when the null hypothesis is true, but rarely occur when the alternative is true. The upshot is that a researcher using our procedure will draw fewer mistaken conclusions than a researcher who just uses statistical significance testing.

To demonstrate, we conduct a simulation to calculate the rate at which the null hypothesis is rejected using two different procedures:

1. statistical significance: reject if two-tailed $p \leq \alpha$.
2. combined statistical significance and substantive robustness: reject if
   (a) result is statistically significant (two-tailed $p \leq \alpha$), *and*
   (b) $c^* > 0$ for $\gamma = 2$.

For the second procedure, we choose a relatively small $\gamma = 2$ in order to ensure a very minimal bar for substantive robustness. We also set the critical threshold for substantive acceptability to $c^* > 0$, the smallest (and easiest-to-pass) threshold for a test of a positive relationship, for the same reason. In short, a result will be judged substantively robust in this simulation if someone with $\gamma \leq 2$ would be willing to act on the alternative hypothesis, even for a very small relationship size.

Our simulation uses a linear DGP:

$$y = 2 + \beta_x x + u.$$

In all simulations, $u \sim \Phi(\mu = 0, \sigma = 0.5)$. We use three values of β_x: $\beta_x = 0.75$ (a "large effect" more easily detected by statistical tests), $\beta_x = 0.25$ (a "small effect" harder to distinguish from noise) and $\beta_x = 0$ (the null hypothesis). For each value of β_x, we conduct 1,000

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

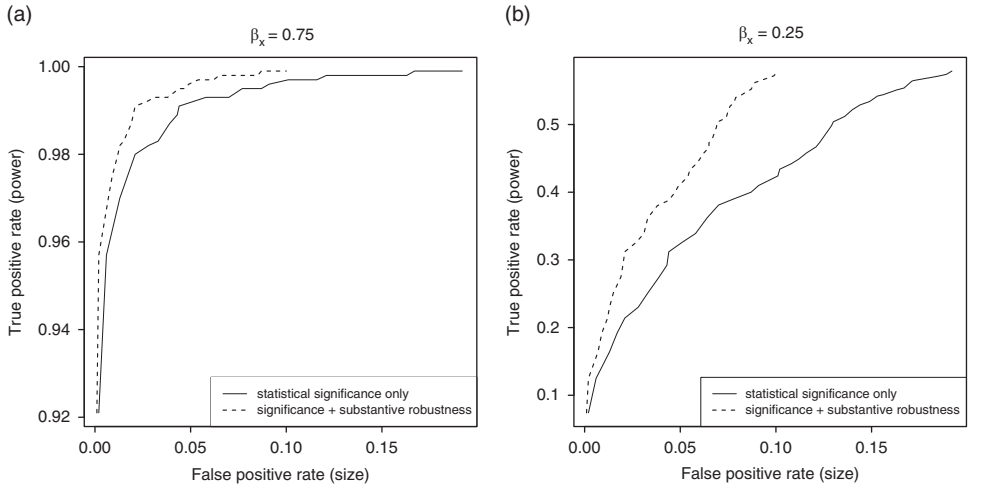


Fig. 4. Size/power analysis for statistical significance testing, with and without substantive robustness checking (a) Large effect (b) Small effect

simulations with datasets of size $n = 100$ each. We then repeat these simulations for values of α (the critical value for the statistical significance test) between 0.005 and 0.2 to get a range of false positive rates. The results are depicted in Figure 4.

Figure 4a plots the proportion of the time that a test procedure rejected the null when $\beta_x = 0$ on the x-axis, and the proportion of the time the test rejected the null when $\beta_x = 0.75$ on the y-axis. Figure 4b plots the same relationship, but with the null rejection rates for $\beta_x = 0.25$ on the y-axis. Ideal performance would mean a completely Γ-shaped curve—an estimator that could achieve a 100 percent true positive rate at a 0 percent false positive rate. Better testing procedures are those that have larger true positive rates for every false positive rate: they can correctly reject false null hypotheses without accidentally rejecting true null hypotheses. As the figure shows, combining statistical significance with a very minimal evaluation of substantive acceptability—a mildly loss-averse researcher ($\gamma = 2$) must be willing to accept an effect of *any* magnitude—improves the power of the test at every size. The performance gap varies according to the nature of the DGP and the α value of the statistical significance test, but combined testing improves power by as much as 10 percentage points in our simulations (for a false positive rate ≈10 percent in Figure 4b).

We also find support for a very useful rule of thumb: if a result is statistically significant but not substantively meaningful ($c^* = 0$ when $\gamma = 2$), in all likelihood the null hypothesis is the correct one. The relevant Bayesian formula for beliefs in this situation is:

$$\Pr(\text{null false} \mid \text{stat. sig., sub. not robust}) = \frac{\Pr(\text{result pattern} \mid \text{null false})\Pr(\text{null false})}{\Pr(\text{result pattern} \mid \text{null false})\Pr(\text{null false}) + \Pr(\text{result pattern} \mid \text{null true})\Pr(\text{null true})}.$$

But in our simulations, there were *no* occurrences in which the result was statistically significant ($\alpha = 0.05$) but substantively not robust ($c^* = 0$ when $\gamma = 2$) when the null was false; *all* of the false negatives in the combined test were cases in which the result was substantively (but not statistically) significant. To ensure the robustness of this finding, we repeat our simulations for $\beta_x = 0.25$ with greater noise in the DGP ($u \sim \Phi(\mu = 0, \sigma = 2.5)$), making false negatives

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

much more likely. We still find that this pattern of results only occurred 1.5 percent of the time under the null hypothesis, leading to an updated belief of:

$$\Pr(\text{null false} \mid \text{stat. sig., sub. not robust}) = \frac{0.015 * 0.5}{0.015 * 0.5 + 0.028 * 0.5} = 0.349,$$

where we determine $\Pr(\text{result pattern} \mid \text{null true}) = 0.028$ by repeating our simulations for $\beta_x = 0$ with $u \sim \Phi(0, 2.5)$. Even under a prior probability of 50 percent that the null is false, the posterior probability that the null is false is only 34.9 percent—nowhere close to conventional significance levels in a hypothesis test. Summarily, our simulation evidence indicates that statistically significant, but substantively non-robust, results are almost surely ascribable to chance.

ASSESSING THE SUBSTANTIVE ROBUSTNESS OF EXISTING RESEARCH

As a technical matter, it is straightforward to assess the substantive robustness of existing research using our technique. We demonstrate using two recently published articles from prominent general interest journals in political science. These examples allow us to show how quantitatively assessing substantive robustness helps refine our interpretation of the results. They also illustrate how our techniques are applied to models with complex marginal effects, such as those from models with interaction (product) terms, where these substantively important quantities are a combination of multiple coefficient estimates and cannot be read directly off of a coefficient table. Finally, we provide documented code for each of these examples, which readers can use to replicate our results or adapt for use in their own inference problems.[12]

Applied Example: Clinton (2006)

First, we will re-examine some of the critical results from Joshua Clinton's 2006 article on representation in Congress in the *Journal of Politics* (Clinton 2006). In this article, Clinton examines the relationship between a survey-based ideology measure of residents of American congressional districts in the year 2000 and legislator ideal points estimated on the basis of voting records.[13] In his ordinary least squares (OLS) regression analysis, he finds that there is a positive relationship between the conservatism of a Republican legislator's voting record and the degree of conservatism expressed by his/her Republican constituents. The same relationship exists between a Republican legislator's record and his/her Democratic constituents' ideology. The conservatism of Democratic legislators' records, by contrast, is associated with the conservatism of their Republican, but *not* Democratic, constituents.

While Clinton uses both OLS regression and an errors-in-variables regression designed to correct for measurement errors, we focus on the OLS results in this replication.[14] We perform the same OLS regression that Clinton ran, the results of which are shown in Table 1. We then calculate a substantive robustness plot for all four effects: the effect of Republican and Democratic constituents on Republican legislators and on Democratic legislators. These plots are shown in Figure 5.

In Clinton's analysis, Republican and Democrat constituents' ideologies have a statistically significant relationship with their Republican legislators' voting records, but only Republican constituents' ideology is related to Democrat legislators' voting records. Substantively

[12] This code is available at http://dx.doi.org/10.7910/DVN/25791.

[13] The legislators ideal points are estimated in Clinton, Jackman and Rivers (2004).

[14] The c^* approach is easily applied to EIV regression, but describing this technique would distract from the central purpose of this article.

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

TABLE 1 *Constituent Influence on "Key Votes" in Congress (Table 3 from Clinton 2006)*

	OLS Rep.		OLS Dem.	
	β	SE	β	SE
Wgt. Same Party Avg. Ideology	1.614*	0.4214	0.1907	0.3290
Wgt. Different Party Avg. Ideology	0.6997*	0.3632	2.167*	0.3241
Constant	0.4602*	0.1182	−.1.145*	0.0919

Note: dependent variable = legislator ideology score from Clinton, Jackman and Rivers (2004). N = 222 (Republicans) and 210 (Democrats). $R^2 = 0.09$ (Republicans) and 0.26 (Democrats). *indicates statistical significance, $\alpha = 0.05$ (one-tailed).

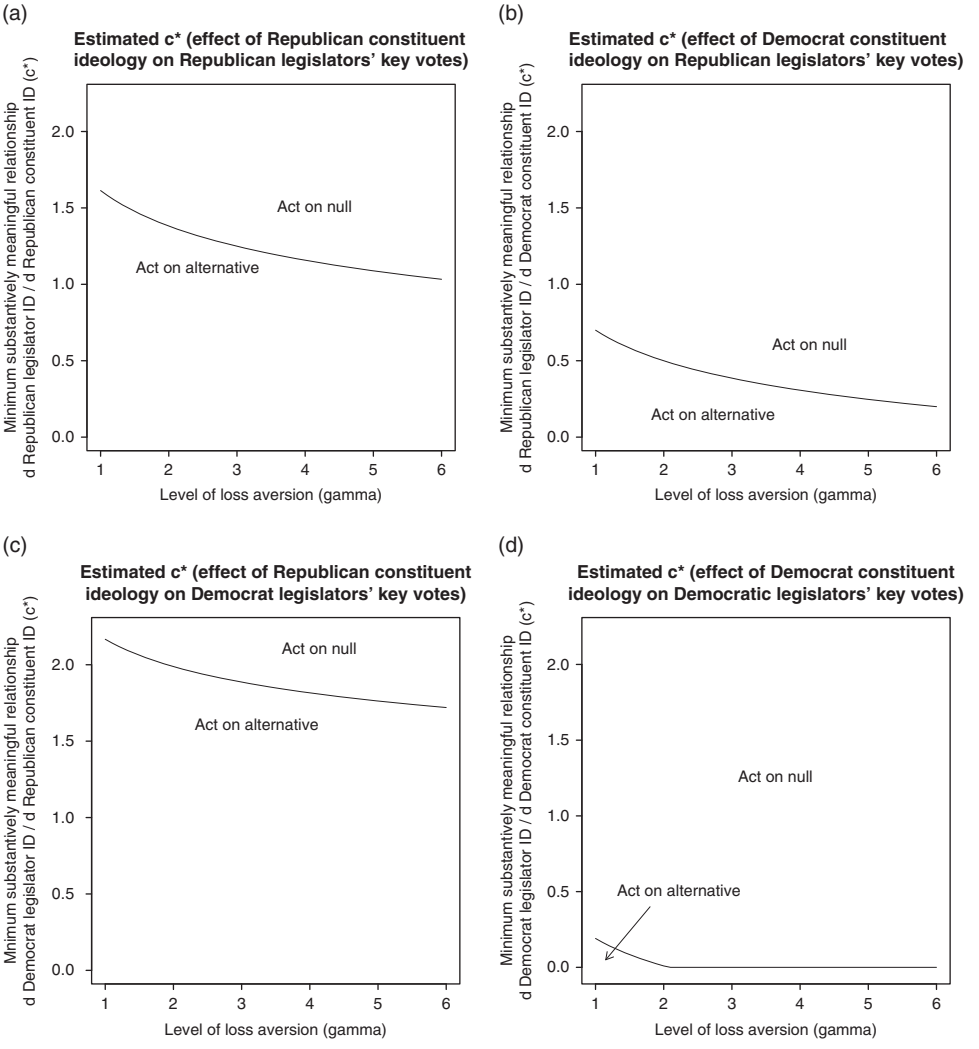


Fig. 5. Substantive robustness assessment for constituent influence on "key votes" in Congress, from Clinton (2006) (a) Republican constituents' influence on Republican legislators (b) Democratic constituents' influence on Republican legislators (c) Republican constituents' influence on Democrat legislators (d) Democratic constituents' influence on Democrat legislators

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

speaking, however, the robustness of these results varies. Republican constituents' ideology seems to have a reasonably robust relationship with both Democrat and Republican legislators' voting records (Figures 5a and 5c): even researchers who value false positives 36 times more than false negatives ($\gamma = 6$) would act on the belief that the coefficient $\gtrsim 1$. With a coefficient of 1, a one standard deviation change in constituent ideology is predicted to move a Democratic legislator at the 50th percentile of his/her party's ideology to the 57th percentile. The effect on Republican legislators is smaller, but still politically significant: a one standard deviation change in Republican constituents' ideology is predicted to move a Republican legislator at the 50th percentile of his/her party's ideology to the 53rd percentile. Visually, we can see that these results are substantively robust, because the range of acceptable coefficients is large across a very wide swath of loss aversion coefficients γ.

The relationship between Democrat constituents' ideology and legislators' voting records is much less robust. This may not be immediately clear from looking at the coefficients in Table 1. For instance, the 90 percent confidence interval for Democratic constituents' effect on Republican legislators is [0.1, 1.3]. This estimate indicates considerable uncertainty about this relationship, but it also firmly excludes the null and includes the possibility of a substantively large relationship. But as Figure 5b shows, a community of heterogeneous researchers would tend to agree on a coefficient of less than or equal to about 0.2—quite close to the left boundary of the 90 percent confidence interval. A coefficient of this size implies that a one standard deviation change in Democrat constituent ideology moves a Republican legislator at the 50th percentile of his/her party's ideology to the 52nd percentile. That is a reasonably small relationship, probably politically ignorable. We conclude that the relationship between Democrat constituents and Republican (or Democratic) legislators is not robust to different standards of substantive judgment. Visually, the non-robustness of these relationships is visible because the range of acceptable coefficient sizes is close to zero (or actually zero) for even modest values of loss aversion γ.

Given our earlier simulation results, which indicated that false positives are minimized by accepting only those results that are substantively robust and statistically significant, we are led to conclude that Clinton's evidence does not support a link between Democrat constituents' ideology and the voting records for legislators of either party. Indeed, when estimating an errors-in-variables regression on the same data, Clinton comes to the same conclusion (but on the basis of statistical significance).

Applied Example: Brambor, Clark and Golder (2006)

In an influential article, Brambor, Clark and Golder (2006, hereafter BCG) describe the importance of interpreting interaction terms substantively, and provide tools and techniques for informative interpretation. These tools and techniques are readily adapted to substantive robustness assessment using our technique. In one example, BCG present a graphical method of presenting marginal effects from linear models with interaction terms. Their example comes from Clark and Golder (2006), a model of the relationship between presidential elections and legislative fragmentation (the number of parties in the legislature). The model is listed in Table 2; note that the relationship between the proximity of presidential elections and the number of electoral parties in the legislature is contingent on the number of presidential candidates.

To test hypotheses about the marginal effect of presidential elections on legislative fragmentation, BCG recommend constructing a plot like the one in Figure 6a, which shows this marginal effect and its 95 percent confidence interval for different numbers of presidential candidates.

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

TABLE 2 *The Impact of Presidential Elections on the Effective Number of Electoral Parties (Table 1 from Brambor, Clark and Golder 2006)*

Regressor	β	SE
Election Proximity	–3.53	0.54
Presidential Candidates	0.33	0.17
Proximity x Pres. Cand.	0.84	0.23
Controls	—	—
Constant	3.12	0.33

Note: OLS model of the number of electoral parties in the legislature. Standard errors are clustered on country. $R^2 = 0.25$, N = 602.

Mathematically, the marginal effect shown in the figure is:

$$\frac{\partial \,\#\,\text{candidates}}{\partial\,\text{presidential elections}} = \beta_{proximity} + (\beta_{prox*cand} * \#\,\text{candidates})$$

and therefore a combination of multiple coefficients from the regression. In Figure 6a, we augment this assessment of statistical significance with a substantive robustness assessment. Specifically, we use our software to determine c^* for each number of presidential candidates using the loss averse utility function from Equation 5 and a variety of different loss aversion coefficients γ.[15] The shaded regions indicate the marginal effect sizes that could be acted upon by researchers with the corresponding level of loss aversion γ. The wider the shaded range, the more robust is the judgment of substantive significance.

As the figure shows, when the number of presidential candidates is small (between 0 and ≈1.5), there is a very robust negative relationship between presidential elections and the number of electoral parties. Even researchers who are extremely averse to false positives ($\gamma > 20$, valuing a false positive 400 times more than a false negative[16]) should conclude that the number of legislative parties participating in an election shrinks by ≈1 when concurrent presidential elections are held (compared to legislative elections held precisely in between presidential terms, such as a US mid-term election). To reduce the interpretation to simple visual terms, there is a broad swath of the marginal effect space that is shaded at a very high level of γ, indicating that even researchers who are extremely averse to false positives would act on the alternative hypothesis of a large marginal effect. In short, researchers with a wide variety of standards for interpreting evidence would accept a large relationship between presidential elections and the number of electoral parties in the legislature.

[15] Inference about the substantive significance of a marginal effect $(dy/dx \mid z) = (y'_x \mid z)$ in an interaction term context uses samples from the distribution $f(y'_x \mid f(\beta \mid data), z)$ to calculate the root in c of:

$$\int \left[u(\text{accept} \mid y'_x(\beta, z), c) - u(\text{reject} \mid y'_x(\beta, z), c)\right]^* f(y'_z \mid f(\beta \mid \text{data}), z) dy'_z = 0 \quad (8)$$

With a sufficient number of samples from $f(y'_z \mid f(\beta \mid \text{data}), z)$ to allow for accurate kernel density estimation of the underlying distribution, the integral on the right-hand side of Equation 8 can be numerically calculated as easily as that of Equation 3 using standard numerical integration packages. Our software allows the calculation of a c^* value when fed a large number of samples from $f(y'_z \mid f(\beta \mid \text{data}), z)$ and a γ value.

[16] To put the preference into perspective, a person with these preferences would refuse to pay $2.50 for a lottery ticket with a 50 percent chance of paying off $1,000!

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

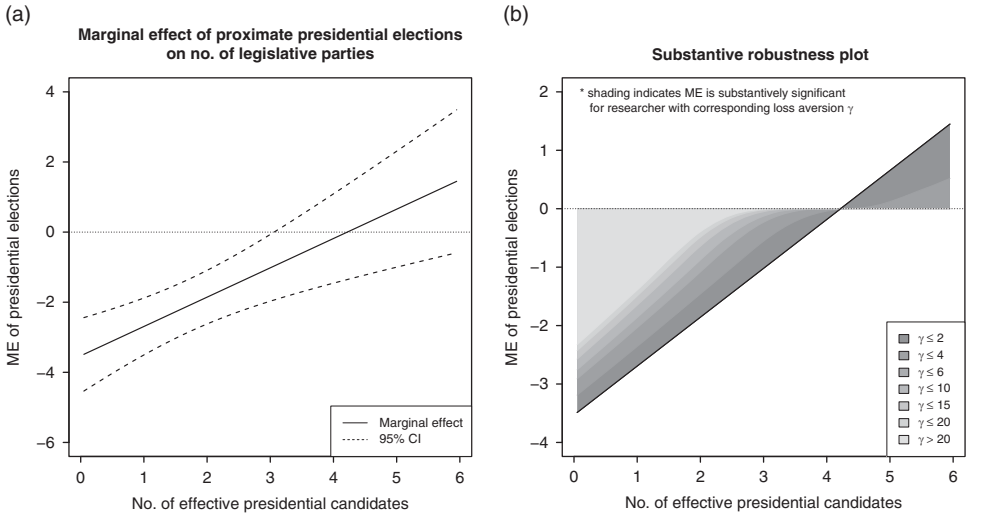


Fig. 6. Robustness assessment for the marginal effect of proximate presidential elections on number of electoral parties in the legislature (Figure 3 from Brambor, Clark and Golder 2006) (a) Statistical significance assessment (b) Substantive robustness assessment

This negative relationship becomes less robust as the number of presidential candidates grows. When there are three presidential candidates, for example, researchers who are even moderately averse to false positives ($\gamma > 4$) would act on the null that there is no relationship at all. Those who are reasonably indifferent between Type I and II errors ($\gamma \leq 2$, valuing false positives 4 times as much as false negatives) would be willing to act on the belief that a reduction of $\approx$1 party occurs when presidential elections are concurrent. These same researchers would also accept that there is a positive relationship between the number of legislative parties and the fact of concurrent presidential elections when there are a very large number of presidential candidates competing for office ($\approx$6). In visual terms, only a small portion of the marginal effect space is shaded, and only for comparatively small γ; this indicates that many researchers would *not* conclude that there is a substantively significant relationship (that is, they would act on the null).

The pattern of substantive findings—a negative relationship between legislative parties and proximate presidential elections when the number of presidential parties is low, and no relationship otherwise—is consistent with the statistical significance findings depicted in Figure 6a, and with Clark and Golder's theoretical hypothesis. Based on our earlier simulations, we therefore have a great deal of confidence in the robustness of this finding.

CONCLUSION

Our article has focused on making three claims. First, substantive significance is easy to subjectively explain (a result strong and certain enough to justify acting upon), but hard to objectively specify because of reasonable differences in scientific judgment. Second, as a consequence of the first claim, it is more productive to ask "would researchers with widely varying preferences consider this result substantively significant?" rather than "is this result substantively significant?" The first question is equivalent to asking, "is this result substantively robust?" Third, requiring that results are substantively robust as well as statistically significant results in more powerful hypothesis tests and fewer erroneous results.

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

It remains for us to argue why researchers should use our particular technique to assess the substantive robustness of their results. It is both possible and useful to think about the substantive significance of results in more heuristic ways. For instance, Achen (1982) recommends examining the bounds of a 95 percent confidence interval of a marginal effect, thinking through the political significance of the relationship if the effect were on either boundary. We certainly do not want to argue against the value of this kind of analysis.

We think our procedure is a valuable supplement to heuristics like those proposed by Achen (1982), because it clarifies the role of heterogeneous researcher standards in the formation of substantive judgments. Applying our technique makes it easier to distinguish between results that are reasonably robust in one way or another—either most people would agree that a relationship exists, or that it does not, based on the available evidence—from those that are more contentious and for which more information is required. From a hypothesis testing perspective, the formality of our procedure also allows us to establish an important benefit of substantive robustness assessment: requiring a relationship to be statistically significant and (minimally) substantively robust improves our ability to distinguish true positives from false positives.

Going a bit further, we also think our procedure is more rational and consistent than heuristic assessment because it is explicitly built on settled, existing work in statistical decision theory and rational choice theory. It ensures that results are objectively evaluated within the reasonable bounds of subjective differences in judgment. And we hope that it is no less convenient than heuristic substantive assessment, because the process is pre-programmed for use in a wide variety of models and can be performed with only a few keystrokes.

Our argument does not hinge on the choice of any particular utility function to be used inside our framework, and indeed we cannot hope to evaluate the infinite space of possible alternatives. We have shown that our particular choice can accommodate widely varying opinions about (1) the proper trade-off between the risks of false positives and false negatives and (2) the minimum coefficient size needed for substantive significance. We have also shown that a statistically significant result that is substantively robust under our particular utility function is extremely unlikely to be a false positive, and that requiring substantive robustness alongside statistical significance improves power (true positive detection) at any fixed size (false positive rate). We believe that is *prima facie* evidence of the value of our choice of utility function; it improves on the status quo of statistical significance testing only (or significance testing with informal arguments about substantive significance). But we must concede that there could be other possibilities that would be even better on these dimensions, and leave this exploration to future work. Our main point is that inferences should be robust to disagreement about important dimensions of judgment, not sensitive to them.

We also believe future work should explore certain applications that would be especially helpful to applied researchers. In particular, not all model products are simple coefficient distributions or even affine functions of coefficients (as in our illustrations). There are, for example, joint F-tests to simultaneously test for the joint statistical significance of a group of coefficients. It would also undoubtedly be helpful to assess the joint substantive significance of a group of coefficients. This would allow a researcher to ask whether a group of marginal effects is collectively influential enough on a dependent variable to be substantively actionable even if no individual effects are.

REFERENCES

Achen, Chris. 1982. *Interpreting and Using Regression. Quantitative Applications in the Social Sciences*. Sage University Press.

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press

Brambor, Thomas, William Clark, and Matt Golder. 2006. 'Understanding Interaction Models: Improving Empirical Analyses'. *Political Analysis* 14:63–82.

Clark, W.R., and Matthew Golder. 2006. 'Rehabilitating Duverger's Theory'. *Comparative Political Studies* 39(6):679–708.

Clinton, Joshua. 2006. 'Representation in Congress: Constituents and Roll Calls in the 106th House'. *Journal of Politics* 68(2):397–409.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. 'The Statistical Analysis of Roll Call Data'. *American Political Science Review* 98:355–70.

DeGroot, Morris. 2004 {1970}. *Optimal Statistical Decisions*. Hoboken, NJ: Wiley Interscience.

French, Simon, and David Rios Insua. 2000. *Statistical Decision Theory*. London and New York: Oxford University Press.

Gelfand, Alan E., and Malay Ghosh. 2000. 'Generalized Linear Models: A Bayesian View'. In *Generalized Linear Models: A Bayesian Perspective*, edited by Sujit K. Ghosh and Bani K. Mallick, 3–22. New York: Marcel Dekker.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis, Second Edition*. Boca Raton, FL: Chapman & Hall/CRC.

Gill, Jeff. 1999. 'The Insignificance of Null Hypothesis Significance Testing'. *Political Research Quarterly* 52(3):647–74.

Hunter, John E. 1997. 'Needed: A Ban on the Significance Test'. *Psychological Science* 8:3–7.

Lunt, Peter. 2004. 'The Significance of the Significance Test Controversy: Comments on 'Size Matters''. *The Journal of Socio-Economics* 33:559–64.

Manski, Charles F. 2007. *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

McCloskey, Deirdre. 1998. *The Rhetoric of Economics*. Ann Arbor: University of Wisconsin Press.

Miller, Jane E. 2008. 'Interpreting the Substantive Significance of Multivariable Regression Coefficients'. In *2008 Proceedings of the American Statistical Association*, Statistical Education Section. Available at http://policy.rutgers.edu/faculty/miller/2008regression_ coefficients.pdf

Pratt, John W., Howard Raiffa, and Robert Schlaifer. 1996. *Introduction to Statistical Decision Theory*. Cambridge, MA: MIT Press.

Siegfried, Tom. 2010. 'Odds Are, It's Wrong: Science Fails to Face the Shortcomings of Statistics'. *Science News* 177:26–29.

Wald, Abraham. 1950. *Statistical Decision Functions*. Oxford: Wiley.

Ziliak, Steven, and Deirdre McCloskey. 2004. 'Size Matters: The Standard Error of Regressions in the American Economic Review'. *The Journal of Socio-Economics* 33:527–46.

——. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.

https://doi.org/10.1017/psrm.2014.14 Published online by Cambridge University Press