Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance

Justin H. Gross University of North Carolina at Chapel Hill

For over a half century, various fields in the behavioral and social sciences have debated the appropriateness of null hypothesis significance testing (NHST) in the presentation and assessment of research results. A long list of criticisms has fueled the so-called significance testing controversy. The conventional NHST framework encourages researchers to devote excessive attention to statistical significance while underemphasizing practical (e.g., scientific, substantive, social, political) significance. I introduce a simple, intuitive approach that grounds testing in subject-area expertise, balancing the dual concerns of detectability and importance. The proposed practical and statistical significance test allows the social scientist to test for real-world significance, taking into account both sampling error and an assessment of what parameter values should be deemed interesting, given theory. The matter of what constitutes practical significance is left in the hands of the researchers themselves, to be debated as a natural component of inference and interpretation.

It may come as a surprise to political scientists just how many empirical methodologists agree with Meehl (1978, 806) that "excessive reliance on significance testing is a poor way of doing science," leading to theories that "lack the cumulative character of scientific knowledge, ...[tending] neither to be refuted nor corroborated, but instead merely fad[ing] away as people lose interest." Meehl's insistence that "the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in [certain areas of psychology] is a terrible mistake, basically unsound, . . . and one of the worst things that ever happened in the history of psychology" (1978, 817) may seem a bit strident, but to the extent that we have come to fetishize the rejection of point null hypotheses, endowing the practice with importance out of proportion to its true value, the spirit of the statement is apt.

Whatever one's philosophy of statistical inference, it is possible—indeed crucial—to do better than reflexively reporting attained thresholds of statistical significance and/or p-values, and then simply regarding these as the focus of subsequent discussion. We have come to expect simple indicators of statistical significance, such as asterisks next to parameter estimates. Such rituals are regularly misconstrued and may even distract from meaningful conversation. While presenting simple summaries of findings may be one worthy goal of lucid scientific communication, neither p-values nor asterisks are suitable for the task. Rather than serving as an invitation to the reader to delve more deeply into potentially meaningful results, secure in the knowledge that apparent patterns are not likely attributable to sampling error, such indicators are often offered and accepted as a *substitute* for interpretation of magnitudes in context.

In what follows, I seek to engage fellow political scientists in a conversation about common practice in the discipline, a conversation that our sibling disciplines (psychology, sociology, economics) have been having for decades but which has largely been absent in political science. In the next section, I briefly review key criticisms of null hypothesis significance testing (NHST). In the third section, I mention some alternatives to NHST that have been suggested by critics of the procedure and propose, as one such alternative, a simple integrated *practical*

Justin H. Gross is Assistant Professor of Political Science, University of North Carolina at Chapel Hill, 361 Hamilton Hall CB#3625, Chapel Hill, NC 27599 (jhgross@unc.edu).

Data from the 1990 ANES Time Series Study may be obtained at http://www.electionstudies.org. I would like to thank Tom Carsey, Frank Baumgartner, Christopher Clark, Terry Sullivan, Thomas Oatley, Heather E. Gross, Justin Esarey, and the participants at the 2011 meeting of the Society for Political Methodology for their feedback on earlier versions of this article. I am indebted as well to Rick K. Wilson, William G. Jacoby, and several anonymous reviewers, who were exceptionally helpful in identifying ways to improve the article.

American Journal of Political Science, Vol. 00, No. 0, xxx 2014, Pp. 1–14

 DOI: 10.1111/ajps.12149

and statistical significance test (PASS-test). This overcomes some troubling shortcomings of NHST, requiring the researcher to assert a range of parameter values that are to be taken as practically or *effectively* null, and thus nudging her toward estimation and interpretation and away from dichotomous pseudo-decision making. Next, I illustrate the procedure with a replication and reanalysis of a media framing study. In the final section, I conclude by suggesting that we pay attention to some recent developments in other fields in acquiring tools that may serve us in more carefully setting up reasonable statistical hypotheses and justifying the conclusions we draw from the data we have.

The technique I propose, hardly a panacea, represents one means of restoring balance to analyses that have dwelled too heavily on existence rather than magnitude. At the very least, I wish to encourage political scientists to join a conversation that has been prominent in the social and behavioral sciences generally and yet nearly absent from political science, beyond a few notable recent exceptions (Esarey and Danneman Forthcoming; Gill 1999; Rainey Forthcoming; Ward, Greenhill, and Bakke 2010).[1] Thoughtful consideration of the status quo will lead us, I have no doubt, to demand more meaningful discussion of results within the studies we consume and those we produce.

The Controversy That Passed Us By: Debating the Merits of NHST

> The function of statistical tests is merely to answer: Is the variation great enough for us to place some confidence in the result; or, contrarily, may the latter be merely a happenstance of the specific sample on which the test was made? The question is interesting, but it is surely *secondary*, auxiliary, to the main question: Does the result show a relationship which is of substantive interest because of its nature and its magnitude? (Kish 1959, 336)

Some time ago, I attended a talk by a scholar doing research on teacher training effectiveness. After explaining his research design, in which teacher success would be operationalized using students' scores on standardized exams, he presented a slide with a long list of coefficients estimated under several models. In keeping with ritual, he called our attention to one or two variables, which had the good fortune to be marked by asterisks. He then noted that, according to his results, parents might do well to ask whether their children's teachers received in-state training; after controlling for a long list of other predictors, the estimated "effect" of in-state training was found to be positive and statistically significant. Asked to interpret model coefficients, the presenter was unable to do so. As it turned out, the expected jump in test scores associated with a teacher being trained in-state was around one-fortieth of a standard deviation! Pressed on whether a parent should seriously be concerned by a (predicted) relationship so small, he conceded that the magnitude did not seem too large, but it was, after all, *statistically* significant. This extreme deference to statistical significance, wherein the very term *statistical* is lorded over the audience as if to imply that it is simply a more rigorous form of everyday significance, leads to opportunities for mischief and—even more perniciously—rewards laziness.

Over the past half century, individual fields in the behavioral, health, and social sciences have grappled publicly with the role significance testing of hypotheses should take in the assessment of research results.[2] The disciplines of psychology, sociology, and economics have devoted entire volumes to the topic (Altman 2004; Harlow, Mulaik, and Steiger 1997; Morrison and Henkel 1970). One of these (Harlow, Mulaik, and Steiger 1997) bears the provocative title *What If There Were No Significance Tests?* This was no empty rhetoric; around the time of the book's publication, the American Psychological Association in fact appointed a task force to consider the recommendation that journal editors ban the reporting of p-values altogether (Wilkinson and Task Force on Statistical Inference 1999). While the proposal did not pass, the task force recommended that estimates of effect size accompany any published p-values, and by 2002, no fewer than 19 journals required the reporting of "effect sizes," a family of standardized measurements meant to identify nontrivial effects of experiments (Thompson 2004).[3]

[1] Carlisle Rainey's excellent contribution (2014) in this journal focuses on the important matter of how to proceed when one's research hypothesis involves a "negligible effect." Our two articles were under development concurrently, address similar themes, and are worth reading as companion pieces.

[2] The history of controversies surrounding the so-called null hypothesis significance testing procedure has typically been traced to passionate disagreements between the towering figures of early twentieth-century statistics, Sir R. A. Fisher on one hand and J. Neyman and E. S. Pearson on the other. A hybrid of Fisher's inferential "significance tests" and Neyman-Pearson decision-oriented "hypothesis tests" would come to be codified in a number of mid-twentieth-century teaching texts, and it is this approach, commonly referred to as "null hypothesis significance testing," that dominates common practice in the social and behavioral sciences. A thorough historical treatment of statistical significance and the roles of Fisher, Neyman and Pearson is provided by Gigerenzer and Swijtink (1990, 79–109).

[3] *Effect size* is a term of art within psychology referring to a number of fully standardized measures (e.g., Cohen's *d*). They tend not

Why all the fuss about p-values? Despite concerns about their widespread misinterpretation—indeed, one's fixation on p-values and asterisks seems to be proportional to one's misunderstanding of what they actually measure—these are merely the most recognizable trappings of an overall framework that overemphasizes minor details. It is not so much their inclusion in analyses that is objectionable as much as their outsized role. As two of the most outspoken critics of NHST assert, "statistical 'significance,' once a tiny part of statistics, has metastasized" (Ziliak and McCloskey 2008, 4), causing many of us to obsess over signal-to-noise ratio in our data—even to the point of forgetting to ask what exactly we are measuring. The more we dwell upon the simple detectability of a signal rather than attempting to characterize and interpret it in context, the more likely we are to be satisfied with the former and permit the absence of the latter.

Numerous authors have presented a long list of criticisms of the NHST approach to social science, and these have been extensively reviewed elsewhere (see, for e.g., Cohen 1994; Gigerenzer 1998; Gill 1999; Meehl 1978; Ziliak and McCloskey 2008). A number of criticisms dwell on persistent misinterpretations of key concepts (p-values, significance levels, null versus alternative hypotheses, the meaning of statistical significance); it is thus tempting to come to the conclusion, as many have, that we need simply to do a better job instructing our students. In fact, some of the most egregious mistakes (e.g., treating p-values as the probability that the null hypothesis is true, or 1-p as the probability of result replication) are banished from our scientific rhetoric during graduate training. The nature of scientific inquiry, though, leads us inexorably to seek within our data the means to assess the relative plausibility of competing hypotheses; however well we might master the acceptable frequentist rhetoric, we cannot help but be unsatisfied by estimating Pr (data|H_0), tantalizingly close as it is to Pr (H_0|data). As Falk and Greenbaum (1995, 94) write, "Significance tests fail to give us the information we need, but they induce the illusion that we have it." Thus, we learn to apply a probabilistic analogue to Aristotle's so-called *modus tollens* syllogism:

> If the null hypothesis were true, these data would be unlikely to have arisen.
> These data have in fact arisen.
> ⇒ The null hypothesis is unlikely to be true.

Although we teach some version of this to our students as the basis of NHST logic, such thinking is not as firmly grounded as it might seem; it has been named the "permanent illusion" and even more provocatively, the "Bayesian Id's wishful thinking" (Gigerenzer 1993, as cited in Cohen 1994). Critics have launched challenges to this extension of deductive logic since at least as long ago as Berkson (1942). It is not so much that the p-value is difficult to define; rather, how we should use the p-value, properly defined, is at issue. A correct interpretation of the p-value simply raises the question of how we should use this information in drawing inferences, from a frequentist perspective. According to the oft-repeated quip of Jeffreys (1961, 385), "What the use of *P* implies is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred."

That key elements of NHST reasoning are so easily and persistently misconstrued in fact reveals a large gap between what researchers expect from hypothesis testing and what NHST actually allows. The two most troubling aspects of the NHST approach in practice are (1) that it compels us to engage in a sort of Kabuki theater, going through the motions of what Rozeboom (1960) has called our "tribal ritual" of rejecting H_0, when we know that with a large enough sample, a point null hypothesis will almost surely be rejected, and (2) the nagging sense that engaging in this empty charade distracts us from delving more deeply into matters of measurement, interpretation, and inference. By repeatedly pretending to be making an up-down decision about a null hypothesis that we know a priori to be false, we risk mindless engagement in this ritualization (Carver 1978; Cohen 1994) and are all too willing to believe that it allows an "automaticity of inference" that "remove[s] the burden of responsibility, the chance of being wrong, the necessity for making inductive inferences, from the shoulders of the investigator and place[s] them on the tests of significance" (Bakan 1966, 430).

Few have bucked the trend and offered spirited defenses of NHST. Frick (1996) emphasizes the point that NHST is especially well suited to ordinal claims, conceding that the approach is not sufficient when the magnitude of a relationship is important. Mogie (2004), while ostensibly writing in defense of NHST, recommends confidence intervals to supplement testing in this manner. Chow (1998, 193), in one of the more enthusiastic defenses of NHST, calls into question the "putative importance of the effect size," claiming that it "is not an index of the evidential support for the substantive hypothesis offered by the data." He concedes that "statistics and practical importance belong to two different domains," but he seems to believe that the two should therefore be segregated.

Unfortunately, we have seen what can result when analysis is constrained to statistical significance without

to be favored within political science and are reasonably eschewed whenever one may meaningfully refer either to native units or some context-driven transformation of these units.

consideration of substantive importance. We wind up placing all emphasis on differentiating signal from noise, leading to such distortions as publication bias, due to the infamous "file-drawer problem" (Rosenthal 1979), the presentation of models that are "only the final tip of an iceberg of dozens if not hundreds of unpublished alternative formulations," rendering estimated standard errors questionable (Schrodt 2006), p-values in published work clustering suspiciously around .05 (Gerber, Green, and Nickerson 2001), and so on—with the likely result that "most claimed research findings are false" (Ioannidis 2005, e124).

With so many drawbacks (just a few of which are listed above), why has this form of significance testing survived and thrived? Clearly, force of habit and the desire for automaticity are difficult to curtail. Yates (1951, 32), blaming a methodological setting of "utmost confusion" at the time of Fisher's major contributions, explains that "in the interpretation of their results research workers in particular badly needed the convenience and the discipline afforded by reliable and easily applied tests of significance." The simplicity and concreteness offered researchers scaffolding on which to build reasonable and reliable habits. Nearly a century after Fisher, we may be ready to let some of that scaffolding fall away in order to discover more flexible approaches to statistical and scientific reasoning.

What Then Might We Do?

Previous Proposals

> When do we stand up and say "Enough already!"? When do we decide that ample arguments have been uttered and sufficient ink spilled for us to stop talking about it and instead start doing something about it? (Levin 1998, 43)

In stark contrast to the scarce attention given the NHST debate in political science, many psychologists have come to regard their discipline as having reached a saturation point, with little to be gained by further rumination over the troubling aspects of null hypothesis significance testing. The above quote comes from J. R. Levin's 1998 article entitled "What If There Were No More Bickering about Statistical Significance Tests?" a not so subtle jibe at the similarly titled book published the previous year (Harlow, Mulaik, and Steiger 1997). The criticisms of NHST had long been piling up, without much serious rebuttal; yet one could detect little change in the practice of scientific communication. Despite near unanimity of sentiment against the dominant practice by those who have spent time thinking and writing about it, no consensus emerged regarding a remedy. The very diversity of proposed options may well have contributed to sustaining the inertia, as it appeared safer (and easier) to just stick with the status quo. Given that one of the key criticisms of NHST is its rigidity and the perceived automaticity with which it is supposed to generate insights, it is unfortunate that the appropriate antidote (a menu of acceptable options to be utilized at the discretion of individual researchers) would fail to attract acceptance. As Cohen (1994 , 1001) warned at the top of his own list of suggestions: "First, don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist." This sentiment, as sensible as ever today, is not new; long ago, Rozeboom (1960, 428) was urging journal editors to "allow the researcher much more latitude in publishing ... statistics in whichever form seems most insightful. In particular, the stranglehold that conventional null-hypothesis significance testing has clamped on publication standards must be broken." Rozeboom's insistence that the individual researcher be permitted to use his "own clinical judgment and methodological conscience [in making] a final appraisal" should be embraced as a guiding principle going forward. Certainly, we should resist the strong temptation to demand a one-size-fits-all set of steps, no matter how apparently sensible any one approach may be in certain situations. When pressed, most scientists will agree with Rozeboom that the "aim of a scientific investigation" is not to reach a binary *decision*, per se, but instead to conduct "a *cognitive* evaluation of propositions." In that case, we should emphasize the *inferential* rather than decision-making model of science in the expository tools we choose.

The best practices of those doing empirical political research today would render the entire debate somewhat irrelevant if more widely adopted; a number of political scientists regularly incorporate the sorts of approaches that NHST critics have long endorsed. Graphical representation of confidence intervals, accompanied by careful interpretations of estimated parameters in the range of plausible values, have become more common, though surprisingly not yet the norm in political science. Bayesian methods (Bakan 1966; Gill 1999), which largely avoid the problems of frequentist significance testing, are increasingly embraced by social scientists. Authors with a sophisticated grasp of statistical methods, and a deep understanding of what these tools can and cannot tell us, find ways to appropriately and imaginatively communicate their results to their audience. Whether or not their publications also happen to include p-values or asterisks alongside estimates is then beside the point.

One suggestion that deserves greater attention is that of Meehl (1978, 817), who urges a more sincere form of falsificationism than that reflected in contemporary caricatures of Popperian procedure, so that theories are genuinely subjected to "grave danger of refutation." Excellent theoretical work on "severe testing" in the philosophy of science literature, most notably by Mayo and Spanos (2006), has connected this notion to lesser known work of Neyman and Pearson. As these authors point out, "a main task for statistical testing is to learn, not just whether H is false, but approximately how far from true H is with respect to parameters in question" (Mayo and Spanos 2006, 329). These authors also seek to move statistical reasoning in science away from a behavioristic (decision-making) philosophy and toward an inferential one. Still, they privilege the testing over estimation mindset. My own approach, outlined below, reverses these priorities so that estimation and interpretation become the primary concern, with the joint "test" of practical and statistical significance serving as a diagnostic aid in this interpretation. The distinction is subtle: In the severity testing approach, while inferences are no longer dichotomous but rather a matter of degree, they still involve degree of support for a claim of existence (rather than of magnitude). Mayo and Spanos insist that confidence intervals fall within the same "error-statistical paradigm" as testing and do not completely avoid comparable problems such as the arbitrariness of the chosen confidence level. However, this is most consequential if one requires a final determination rather than discussion of plausible values within a range of uncertainty. "Although CI's can be used ... as surrogates for tests, the result is still too dichotomous to get around fallacies: it is still just a matter of whether a parameter value is inside the interval (in which case we accept it) or outside it (in which case we reject it)" (Mayo and Spanos 2006, 347). Here, the up-down decision-making perspective persists; indeed, when a confidence interval is considered from a pure testing perspective, it is equivalent to a corresponding test. However, as we shall see in the examples below, confidence intervals may also be considered on their own terms.

When conducting hypothesis tests, what researchers typically have in mind is not a literal interpretation of the null hypothesis as a single point hypothesis, but rather that "the value of μ is close to some specified value μ_0 against the alternative hypothesis that μ is not close to μ_0" (DeGroot and Schervish 2002, 481). It makes more sense to replace the idealized simple hypothesis with "a more realistic composite null hypothesis, which specifies that μ lies in an explicit interval around the value μ_0" (482). It is this suggestion that I take as the basis for synthesizing statistical and substantive significance within one procedure. Serlin and Lapsley (1985) suggest a similar approach, calling this interval a "good-enough band" around the point null value.[4]

One may reasonably protest that such a procedure requires an arbitrary choice of length of the interval constituting a composite null set. DeGroot and Schervish (2002, 519) recommend a posterior probability plot for different values of the null interval diameter δ, but no such plot is available to the frequentist. It is nonetheless possible to conduct analyses of sensitivity to the choice of δ (as well as the choice of confidence level $1 - \alpha$), as a robustness check. In any case, explicit discussion of what difference or relationship would be meaningful is an essential, but frequently overlooked, task for the researcher. As DeGroot and Schervish (2002, 520) write, "forcing experimenters to think about what counts as a meaningful difference is a good idea. Testing the (simple) hypothesis ... at a fixed level, such as 0.05, does not require anyone to think about what counts as a meaningful difference." This observation cuts straight to the heart of why the conventional NHST approach encourages bad habits. Indeed, demanding that political scientists articulate and even debate what would constitute a meaningful effect in context of their particular research problems, rather than encouraging arbitrary cut points, puts the emphasis back on experts' subject-area knowledge; political scientists should welcome the opportunity rather than shrink from it.

Practical and Statistical Significance: The PASS-Test Integrated Approach to Detecting Meaningful Magnitudes

Given parameters of substantive interest, be they real-world quantities (e.g., difference between mean incomes for two subpopulations) or quantities whose meaning is derived only within a proposed model (e.g., a Poisson regression coefficient), a researcher wishing to simultaneously test for statistical and substantive significance should begin by declaring a set of parameter values to be taken as *effectively null*. This should be based, when feasible, upon context and defended by the author. Such a choice should emerge from reflection on the following question: If one could know precisely the "true" value of a parameter, what values would seem inconsequential and which would seem worthy of note? The resulting null set

[4] Placing emphasis on traditional testing properties such as control over Type I error, based implicitly upon the assumption of a 0-1 loss function, leads to reliance upon noncentral F and t distributions. See also Steiger and Fouladi (1997) for recommendations when taking this approach.

would include any value that seems *practically indistinguishable* from the (sharp) null value, or *effectively null*. The very process of thinking this through and the resulting conversation with others would itself be a healthy development. Indeed, while the precise distinction between what values are effectively null and which ones are of interest may be somewhat arbitrary, thoughtful consideration should in most cases reveal a range of values that all knowledgeable individuals would take to be effectively null and a range of values that anyone would consider noteworthy. In the next section, I will illustrate how one may go about proposing such a null set, as well as the usefulness of conducting a simple sensitivity analysis in order to indicate how robust one's results are to both this partition of the parameter space and the chosen level of confidence/statistical significance.

The effective null set may be used as a heuristic for the reader who wishes to get a quick sense of what the authors purport to be of value in their results. Suppose, for example, we wish to know whether two groups of laborers, comparable other than with respect to gender, earn the same hourly wage, on average. We are unlikely to care if the true difference is only a few cents, even if this difference were known with absolute certainty. Suppose we declare the effective null set to be $\Theta_0 = [-\$0.25, \$0.25]$, so that any discrepancy of 25 cents or less is considered inconsequential or insufficiently notable to merit intervention. Then once a confidence interval is constructed from the data (at the preferred level, say 95%), simple qualitative distinctions may be drawn:

1. The confidence interval may lie entirely *outside* the effective null set, in which case we may say with 95% confidence that we have detected a *meaningful difference* (i.e., one that is *practically/substantively significant*).
2. The confidence interval may lie entirely *within* the effective null set, in which case we may say with 95% confidence that there is *no practical difference.*
3. The confidence interval may *overlap* the effective null set, in which case we may say that the test of practical and statistical significance is *inconclusive* at 95% confidence. Greater precision is needed in order to disambiguate the results.

This represents an oversimplification of results, but at least an oversimplification that points in the direction of what the reader should care about. If the confidence interval lies mostly within the null set, we might say the evidence "leans against" a meaningful difference; if barely overlapping Θ_0, we might say the difference is "likely" meaningful. Note that this notion of what is to be taken as meaningful incorporates *both* statistical and substantive significance. Having an agreed-upon shorthand that may draw the casual reader to closer inspection could be helpful. In the PASS-test presentation, the simplification addresses what is of scientific interest (the magnitude of a parameter) while simultaneously providing evidence as to whether we can have some confidence that a seemingly meaningful result is not a phantom. Not inconsequentially, the "test" requires presentation of confidence intervals and promotes careful thinking about the set of plausible parameters. In this way, it is a test that actually weans us from excessive dependence on the testing paradigm itself.

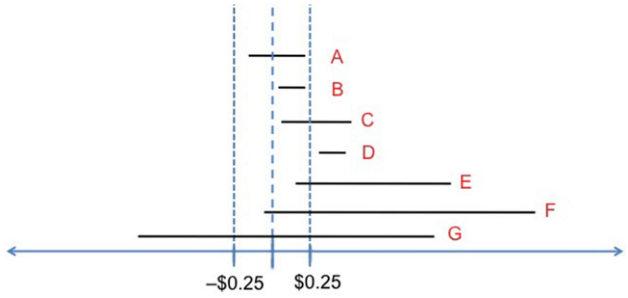
Two especially troubling aspects of NHST are addressed by the combined practical and statistical significance testing framework outlined below. First, by never requiring a point null hypothesis to compete with a composite (interval) research hypothesis, one abandons the aforementioned practice of using H_0 as a straw man that is known to be false before data are even examined. Instead, it will be at least hypothetically possible to legitimately find support for the null hypothesis. As n gets large, the width of the resulting confidence interval will shrink until it lies entirely in either the effective null interval or the alternative. Furthermore, the often disingenuous proposal of one-sided hypotheses, of the form $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$, may be replaced by the consideration of genuinely commensurable parameter sets: $H_0 : \theta < \theta_0$ vs. $H_1 : \theta \geq \theta_0$. The use of one-tailed tests has itself long been the subject of controversy (see, e.g., Eysenck 1960), in part because of the suspicion that they are employed more out of a desire to compensate for poor power to reject the null than out of any theoretical motivation. In principle, the frequentist test of point null versus one-sided interval acts as a proxy for the two-interval test we really have in mind; the point null is chosen to have the sampling distribution most likely to generate a test statistic in Θ_1, the set of parameters consistent with the research hypothesis H_1. Bayesian analysis offers the most natural way to allow competition between two interval hypotheses, as one may simply integrate the posterior distribution over each subset of the parameter set and compare directly or consider the corresponding odds ratio. In lieu of the Bayesian alternative, inspection of the confidence interval relative to the partitioned parameter space, with the border between the two sets chosen so that only nontrivial values lie in Θ_1, would seem to be more informative than the usual implementation.

In Figure 1, I plot confidence intervals for hypothetical results one might obtain in addressing the wage difference question. Since the results are imagined, let's suppose for the sake of concreteness, that they are 95%

TABLE 1 "Decisions" in PASS versus NHST Applied to Hypothetical Results in Figure 1

	PASS-Test	NHST
A.	No meaningful difference (H_0 supported)	Do not reject H_0 (Inconclusive)
B.	No meaningful difference (H_0 supported)	Find evidence of a difference (H_0 rejected)
C.	Possibly trivial difference (Inconclusive)	Find evidence of a difference (H_0 rejected)
D.	Find meaningful difference (H_1 supported)	Difference distinguishable from 0 (H_0 rejected)
E.	Trivial, small, or large difference (Inconclusive)	Find evidence of a difference (H_0 rejected)
F.	Wide range of null/+ values (Inconclusive)	Do not reject H_0 (Inconclusive)
G.	Wide range of –/null/+ values (Inconclusive)	Do not reject H_0 (Inconclusive)

FIGURE 1 Seven Hypothetical Confidence Intervals to Be Interpreted



Note: The effective null set is bounded by the outer dashed lines; a difference of up to 25 cents per hour has been deemed inconsequential. Interpretations are offered in Table 1, along with each corresponding interpretation under conventional NHST.

confidence intervals (one could also superimpose two or three confidence intervals with different α's). I have indicated with dashed segments the thresholds of my effective null set and $0 as the corresponding sharp null from a conventional analysis. For each interval, consider how results would be interpreted under the proposed PASS-test in contrast to how the corresponding NHST would be interpreted (Table 1). In neither case should the researcher be satisfied with simply reporting simple direction and significance. Such a simplified summary of results is useful, but it cannot replace a discussion that considers the range of plausible values falling in a confidence interval and interpreting their meaning.

Stated precisely, in algorithmic form, a combined practical and statistical significance test includes the following steps, assuming a continuous parameter space for a parameter reflecting some relationship of interest:

Practical and Statistical Significance Test (PASS-Test)

1. Partition the parameter space Θ for each estimate of interest into Θ_0, an *effective null set*, and Θ_1, a set of *contextually meaningful values*, both noncountable sets corresponding to composite hypotheses.
2. Defend the choices of each partition either through existing theory or reference to other covariates.
3. Estimate parameters using confidence intervals.
4. For an estimated $1-\alpha$ confidence interval C,
   (a) find in favor of the null hypothesis of no meaningful relationship if $C \subset \Theta_0$, with $1-\alpha$ confidence;
   (b) find in favor of the alternative hypothesis of a meaningful relationship if $C \subset \Theta_1$, with $1-\alpha$ confidence;
   (c) or declare the result inconclusive at $1-\alpha$ if $C \cap \Theta_0 \neq \emptyset$ and $C \cap \Theta_1 \neq \emptyset$, that is, if the confidence interval overlaps the two sets of values.
5. Employ sensitivity analysis to determine whether other reasonable partitions of Θ or choices of α would have affected the outcome of the test.
6. Discuss and interpret the confidence intervals in context, noting the range of likely effect sizes. In particular, if the result must be declared inconclusive at the selected α, the analyst should, for example, distinguish between a fairly precise confidence interval containing parameter values either in or close to the effective null set and one that is wide (less precise), but containing mostly values considered to be substantively significant and perhaps even large. In the latter case (e.g., E and F in Figure 1), the collection of additional data, if possible, should be recommended so that increased precision may help researchers distinguish a non-null relationship.

An Illustration: Media Effects on Public Opinion

> A major problem involved in adjudicating the scientific significance of differences is that we often deal with units of measurement we do not know how to interpret. (Carver 1978, 389)

I next illustrate how the recommended approach may serve as one way to enrich the presentation and interpretation of results. In certain instances, the PASS framework for considering results may strengthen authors' arguments; in other cases, it makes more obvious the tentativeness with which the results should be viewed.

In their article exploring the three most widely studied types of media effects (agenda setting, priming, and framing) in the lead-up to the Persian Gulf War of 1990–91, Iyengar and Simon (1993) consider whether exposure to television news may predict support for a military response to Iraq's invasion of Kuwait and the subsequent crisis. Having studied eight months of primetime newscasts during the relevant time interval, they note the predominant use of *episodic* over *thematic* framing; based on extant theory and previous research, they suspect that this will lead to viewers' attribution of responsibility to particular individuals and groups rather than broader historical, societal, or structural causes, and they anticipate that this will translate into support for the use of military force against Saddam Hussein rather than diplomatic strategies by those who consume such media. They regress a variable measuring respondent support for a military over diplomatic response on several predictors, including presumed indicators of news exposure. The measurements are taken from the 1991 ANES Pilot Study (Miller et al. 1999).

According to the logic of combined practical and statistical significance testing, it is essential that one consider what magnitudes of coefficients would be impressive *if one were able to observe the parameter values themselves without sampling error*. Iyengar and Simon (1993) are primarily concerned with the expected effect on military support corresponding to variation in the values of *TV News Exposure* and *Information*, measures of, respectively, television news consumption and awareness of political information via identification of political figures in the news. Conditioning on party, gender, race, education, and general support of defense spending, what sort of coefficients should we view as *effectively zero*, and, conversely, what values would indicate at least a somewhat meaningful relationship? This sort of question, as previously noted, is too often left unasked. To the extent that it does arise, it is almost always handled completely informally. To their credit, the authors here distinguish between the two types of significance: "Overall, then, there were statistically significant traces of the expected relationship. Exposure to episodic news programming strengthened, albeit modestly, support for a military resolution of the crisis" (Iyengar and Simon 1993). From a PASS-test—rather than NHST—perspective, the assessment of the degree to which this type of programming corresponds to greater military support is of principal concern; the claim of a "modest" relationship should therefore be more carefully explained and supported.

The Iyengar–Simon model may be written in the following manner:

$$\begin{aligned} &E(MilitarySupport) = \\ &\beta_0 + \beta_1 TVNews + \beta_2 Knowledge + \beta_3 Male \\ &+\beta_4 nonWhite + \beta_5 (Male \times Knowledge) \\ &+\beta_6 (nonWhite \times Knowledge) + \beta_7 Republican \\ &+\beta_8 DefenseSpend + \beta_9 Education, \end{aligned} \tag{1}$$

with departures from conditional expectation assumed distributed $\in \sim$ Normal$(0, \sigma^2)$. The predictors of primary interest are $TVNews$, the number of self-reported days per week watching TV news, and $Knowledge$ (called $Information$ in the original article), the respondent's score from 0 to 7 on a quiz of recognition of political figures, taken as another proxy for news consumption. In Iyengar and Simon's (1993) model, the contribution of $Knowledge$—but not $TVNews$—is allowed to vary by race and gender (through the inclusion of interaction effects), so that, in addition to the coefficient on $TVNews$, one should wish to learn whether the following parameters are of a meaningful magnitude:

$$\begin{aligned} \gamma_1 &= \beta_2 = \text{effect of } Knowledge \text{ among white females} \\ \gamma_2 &= \beta_2 + \beta_5 = \text{effect of } Knowledge \text{ among white males} \\ \gamma_3 &= \beta_2 + \beta_6 \\ &= \text{effect of } Knowledge \text{ among nonwhite females} \\ \gamma_4 &= \beta_2 + \beta_5 + \beta_6 \\ &= \text{effect of } Knowledge \text{ among nonwhite males} \end{aligned} \tag{2}$$

Within each demographic category, the associated composite parameter is interpretable as the difference in expected level of support for a military solution associated with an additional point on the political knowledge quiz. Thus, for example, if $\gamma_1 = 0.25$, this means that one might expect an extra correct answer on the quiz taken by a white

female to correspond to an additional quarter point on the intervention support scale ranging from 1 to 4, assuming that the ordinal scale of support for diplomacy versus military action can be sensibly interpreted as if it were an interval-level measurement. A large difference of 4 points on the quiz (e.g., correctly identifying six rather than two political figures, or four rather than zero) would be expected to translate into a full extra unit in the support for a militaristic solution on the scale ranging from 0 to 4.[5] Understanding this allows the researcher to set up reasonable expectations of what might be considered a truly meaningful "effect"[6] and then evaluate whether the data support such a finding in light of sampling error.

Following the steps outlined above, one would begin by declaring a reasonable null set. Here is an opportunity for the researcher to utilize his or her applied knowledge. Two options for handling this are as follows:

- *Analyze the results on their own terms.* This is probably the best route when theory is well developed, and the researcher is extremely comfortable with the data, its units of measurement, and the modeling technique employed, and wishes to engage other experts in a discussion of what magnitudes should be considered meaningful.
- *Establish benchmarks.* This may be accomplished by anchoring discussion of key predictors to other covariates believed a priori—and confirmed empirically—to be large. In this example, one's party and support of defense spending should provide good benchmarks, as they will be expected to be strong predictors of support for the use of military force. If predictors are in comparable units (e.g., all binary), proposal of the effective null set may be conducted with reference to these units. Another option, which I illustrate, is to standardize the predictor variables only; thus, one may talk about the expected difference in response corresponding to a large, moderate, or small change in each predictor, defined in terms of that predictor's variability.

[5]Specifically, 1 point was awarded if the respondent supported tougher military action going forward, rather than any of three less hawkish alternatives, and from 1 to 3 points were awarded for level of militarism expressed in response to the question of what the United States *should have done* as an original response to the Persian Gulf crisis.

[6]Whether or not the association estimated via regression on the observational data is in fact causal is peripheral to the current discussion; for the sake of discussion, we will assume any observed association to be causal.

TABLE 2 Replication of OLS Results on Exposure to Information/TV News as Predictors of Support for Military Response in Gulf (Iyengar and Simon 1993)

Support for Military Rather Than Diplomatic Response			
	b	***SE***	***p* level**
TV news exposure	0.017	0.010	.105
Knowledge	0.079	0.031	.012
Male × Knowledge	−0.079	0.039	.039
Nonwhite × Knowledge	0.064	0.052	.221
Male	0.623	0.101	<.001
Nonwhite	−0.659	0.114	<.001
Republican	0.077	0.013	<.001
Defense spending (favor)	0.191	0.018	<.001
Education	0.072	0.017	<.001
Intercept	0.984	0.126	<.001
Adjusted R^2	0.19		
n	1,778		

The table of results found in the original article closely matches the results from my own replication,[7] shown here in Table 2. This table reflects typical NHST-style presentation and thus is not, on its own, especially informative, and so the p-values attract the bulk of the reader's attention. The authors discuss these results somewhat vaguely:

> Partisanship, race, gender, and education—all affected respondents' policy preferences concerning resolution of the conflict. Republicans, males, those with more education, and Whites tended to support the military option. Support for increased defense spending was *strongly* associated with a more militaristic outlook toward the conflict. Both indicators of exposure to television news exerted significant effects—more informed respondents and respondents who watched the news more frequently were most apt to favor a military resolution. The effects of information were *markedly stronger* among women and minorities. . . . (Iyengar and Simon 1993, 380, emphasis added)

As is the norm, the assessments here are mostly all-or-nothing. Predictors either "affected" outcomes or not, and when the strength of a relationship does get mentioned, the basis on which this judgment of magnitude

[7]The authors clearly explained their construction of variables from the ANES data; slight discrepancies likely reflect minor ambiguities in the description, but coefficient values and standard error estimates are nearly identical.

(e.g., "strong," "markedly stronger") is made is unclear. Moreover, not having to consider or communicate what would constitute a minimally meaningful result makes it easier to avoid interpretation of the interaction effects in the current example, although these stand to shed the greatest light on the central empirical question here.

Using the benchmark approach discussed above, consider a possible PASS analysis based on the estimates of coefficients and standard errors.[8] The 95% confidence interval for support of defense spending is (0.15, 0.23), but as a simple point of comparison, consider an estimate of around 0.20. The variable measures response to whether the nation should decrease or increase defense spending on a scale ranging from *greatly decrease* (1) to *greatly increase* (7). The median, and by far the modal, response is support for the status quo; 43% of respondents preferred neither an increase nor a decrease in defense spending. $Q_1 = 3$ and $Q_3 = 4$, so the interquartile range (IQR) $= 1$. There is little variation in responses to this question, so that even a single unit (level) difference is relatively large and provides a key benchmark for comparison with predictors of interest. One would not expect that a more subtle predictor such as TV news watching would possibly match the predictive power of the respondent's general predisposition toward a muscular defense, but it would be tough to argue that a magnitude absolutely dwarfed by this benchmark variable would still be noteworthy. It should not be controversial to consider as effectively zero anything less than a 0.01 expected difference in response level per predictor IQR. Recall that the index of support for a military strategy in Iraq ranges from 1 to 4, so in absolute terms, this means that anything less than one-hundredth of a full-level jump in response per large change in predictor value, conditioning on the other explanatory variables, is to be considered negligible. In relative terms, no less than one-twentieth of the per-IQR explanatory power of *Defense*, or around two-fifths that of *Republican* party identification, will be deemed meaningful.

Consider what is communicated in Figure 2, as opposed to the NHST-based Table 2. Aside from *Defense*, the baseline used to anchor our judgment of effect size, where can we distinguish meaningful associations? The authors assert a "markedly stronger" effect of information on minorities and women, but is this so? Looking at the results, no such relationship is detected for nonwhite men. Information makes a meaningful difference specifically for some women, it would seem. For nonwhite women, in particular, greater information exposure clearly corresponds to higher expected level of support for military intervention, all else equal. For white women, the relationship likely exists as well; the per-IQR point estimate is comparable to that for party identification, *Republican*, and *Education*, with a 95% confidence interval entirely outside the null interval. In both cases, the association may actually be quite large; a sample with more women would increase the estimate precision and allow us to find out. Controlling for the measure of political knowledge, as well as the other variables, exposure to television news is not predictive of the dependent variable at a level distinguishable from the identified set of null values. In fact, a two-way table reveals no discernible marginal relationship between self-reported TV news exposure and hawkishness with regard to the Gulf crisis.[9] With little evidence of a relationship even before conditioning on covariates and relying on linearity assumptions, it is not surprising no more than a tiny association is detectable according to the regression results presented in Figure 2.

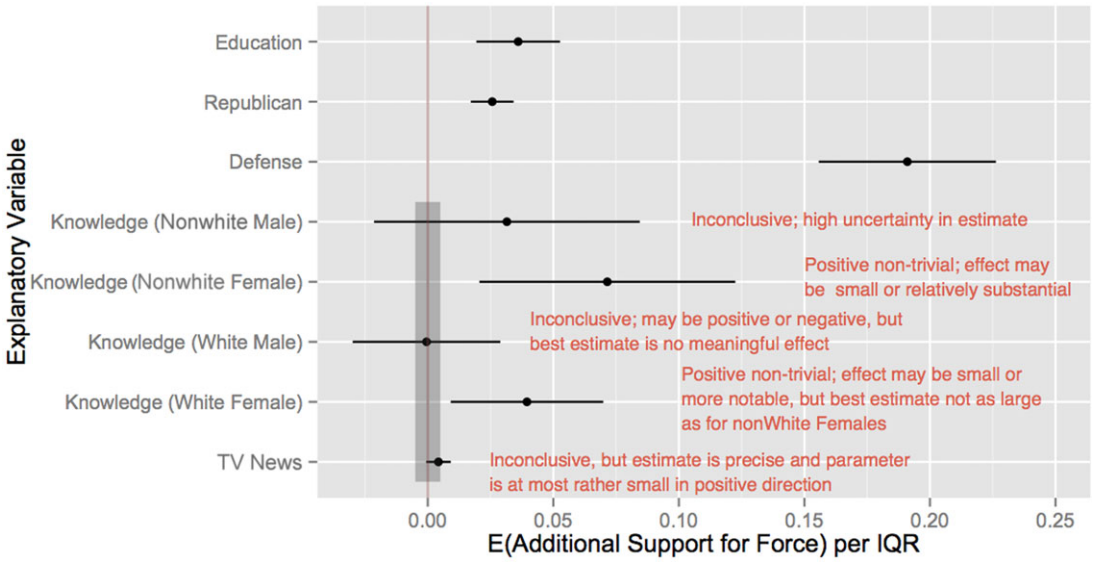
As others have pointed out, confidence intervals rely upon a seemingly arbitrary choice of confidence level; similarly, the width of the effective null appears subjective, if not exactly arbitrary. Thus, one should indicate how sensitive one's claims are to the choice of these two quantities. In Figure 3, three confidence levels are displayed, and a more conservative $[-.03, .03]$ effective null set is superimposed upon the initially proposed $[-.01, .01]$ set. An examination reveals how robust the results are to these two choices. For example, using the wider null set, any reasonable confidence interval for TV news exposure indicates a substantively insignificant relationship with the response variable. For white males, the null hypothesis of no meaningful relationship between political knowledge militarism is *accepted* at the 67% confidence level. On the other hand, this null hypothesis is rejected in the case of nonwhite females at both 67% and 95%, and white females at 67%. Sticking with the original narrow null set, the PASS-test for knowledge versus militarism, controlling for the rest, is inconclusive for all four demographics at the more rigorous 99% level.

Finally, in considering the presentation of results in Figures 2 and 3, some may worry that the analysis relies too heavily on verbal description. I happen to find verbal interpretation a strength, rather than a weakness, but note that all useful information contained in conventional tabular presentation is contained in the figures as well. The

[8] Approximate standard errors for composite parameters are calculated as $\hat{se}(\hat{\gamma}_2) = \sqrt{\widehat{Var}(\hat{\beta}_2) + \widehat{Var}(\hat{\beta}_5) + 2\widehat{Cov}(\hat{\beta}_2, \hat{\beta}_5)}$, $\hat{se}(\hat{\gamma}_3) = \sqrt{\widehat{Var}(\hat{\beta}_2) + \widehat{Var}(\hat{\beta}_6) + 2\widehat{Cov}(\hat{\beta}_2, \hat{\beta}_6)}$, and $\hat{se}(\hat{\gamma}_4) = \sqrt{\widehat{Var}(\hat{\beta}_2) + \widehat{Var}(\hat{\beta}_5) + \widehat{Var}(\hat{\beta}_6) + 2[\widehat{Cov}(\hat{\beta}_2, \hat{\beta}_5) + \widehat{Cov}(\hat{\beta}_2, \hat{\beta}_6) + \widehat{Cov}(\hat{\beta}_5, \hat{\beta}_6)]}$.

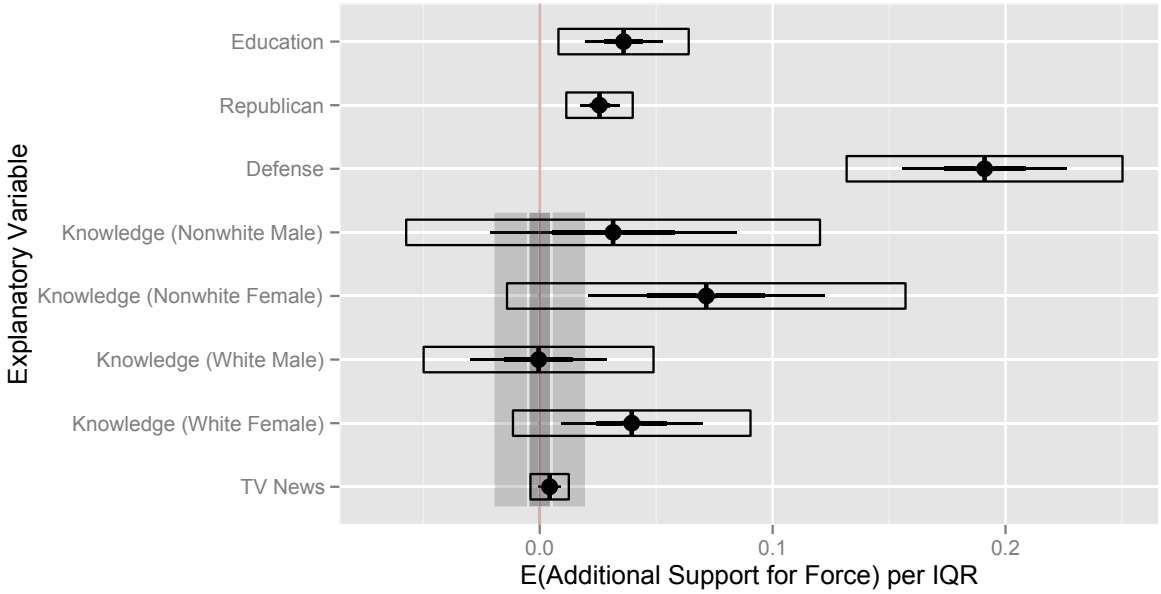
[9] Perhaps surprisingly, over 50% of respondents claim to have watched television news every day of the previous week. According to Prior (2009), news watching is vastly overreported, so this may be a quite noisy measure.

FIGURE 2 Estimates of Parameters of Particular Interest from Equations (1) and (2)



Note: Estimates are expressed as 95% confidence intervals on the rate of change in respondent's support for military strategy in Iraq per interquartile range increase in each explanatory variable; baseline covariates appear at the top and an effective null set appears in gray.

FIGURE 3 Sensitivity Analysis



Note: Results from Figure 2, this time presented with both a conservative and liberal plausible null set, as well as three values for confidence level $1 - \alpha$. The short segments indicate 67% confidence intervals, the longer segments are the 95% intervals, and the large rectangles depict 99% confidence intervals.

explanations superimposed on Figure 2 are for pedagogical purposes only and need not be an expected feature of such a graph. I have found there to be clear consensus that it is poor practice to simply instruct the reader to "see results in the table provided," without offering the author's own interpretation. This is true whether one summarizes results in a conventional table or in a figure depicting confidence intervals or posterior distributions. It is incumbent upon the author to refer to tables, figures, and other quantitative and qualitative evidence in making

his or her argument. It is then up to the engaged reader to consider the evidence presented, alongside previous work, and determine whether the author's particular interpretations are convincing. There is no denying that interpretation of graphs involves subjectivity. The avoidance of this subjectivity can only be bought through the acceptance of automaticity, taking us again down a path that has not served social science well. Our instinct to seek a one-size-fits-all criterion for practical significance is understandable, but it is even less tenable than such a standard for statistical significance.

One might suspect that this subjectivity will be an invitation to choose one's null set solely to obtain the illusion of significance, but this is already the case with the status quo, as we have seen. Similarly, the potential for meta-analysis of multiple related studies may seem to be hindered by the lack of uniformity among investigations that employ different null sets, but of course, studies purporting to examine the same phenomena have always differed in all sorts of ways. In fact, while variability in the choice of measure, model specification—and now the effective null set—requires extra time and care on the part of the meta-analyst, this heterogeneity of study detail offers its own advantage: Apparent consensus will not be attributable a particular arbitrary choice, repeated reflexively by each subsequent researcher.

Overreliance on arbitrary but agreed-upon thresholds simply leads to consensual mass manipulation, in which passing a magical threshold ends rather than begins the conversation. Just as there is no formula that generates theories for us, there is no statistical technique that can produce objective standards of what facts should be scientifically impressive. The job of convincing a critical audience that a certain state of the world would be interesting, surprising, or even shocking—if true—is entirely our own, whether done so informally, by suggesting an effective null set, or by continuing to pass off statistical distinguishability alone as sufficient. This essential task of the social scientist should be embraced; to do otherwise is an abrogation of our responsibility as scholars and an underestimation of the capacity of our audience. Rather than supposing that a particular effective null set is somehow acceptable to all, it would be preferable to use sensitivity analysis to investigate the robustness of one's results to other defensible null sets (and confidence levels), as I have done briefly in this illustration. If readers find our chosen null set to be suspiciously convenient, or our consideration of alternative null values overly restrictive, they should challenge us and debate the point.

Conclusions

The criticisms of null hypothesis significance testing as commonly put into practice are widely recognized and not especially controversial among statisticians. Despite having been aired in numerous forums over the decades—at least outside of political science—not much has changed in terms of standard practice. Bayesian approaches to hypothesis testing, focusing attention directly on the relative probabilities of competing hypotheses in light of data, do not suffer the principal failings of the NHST, and so the growing acceptance of the Bayesian toolkit has been a source of improvement on this front. Within a frequentist framework, however, debates over the merits of significance testing have had little impact on practice. It is difficult to change habits while the status quo is rewarded and no consensus nor even clear guidance on a preferable alternative emerges. In employing non-Bayesian forms of analysis, it would be wise to develop a habit of summarizing the results of parametric empirical analyses in a manner that emphasizes *statistical* distinguishability from a practically inconsequential set of parameter values. The expectation that scholars consider *magnitudes* of parameters of interest, and not be content to simply distinguish signal from noise, requires little additional work and no special training, but guides us in the direction of meaningful discussion and away from the seductive lure of empty formalism. Additionally, to convincingly argue about what results should be deemed significant in practical terms provides incentive for creative intertwining of qualitative with quantitative knowledge of subject matter.

In addition to encouraging fellow political scientists to join our colleagues throughout the social sciences in reflecting upon potential improvements to common implementation of NHST, I would be gratified to see others give thought specifically to how we might choose sets of effectively null parameter values when working with more complicated models. It remains to be seen whether the PASS-test might find fruitful application in fitting, say, generalized linear models, exponential random graph models, and others from among various families of sophisticated parametric models that have been proposed for use in various applications. And yet, even the *attempt* to bring the question of practical significance into our methodological decision-making process would be a worthwhile exercise. Social scientists may recall a time, as logistic regression was just beginning to gain popularity in our disciplines, when it was common to simply avoid any interpretation of parameter estimates whatsoever, usually with some sort of disclaimer that the model coefficients were not easily interpretable. Fortunately, this is no longer

typically tolerated; we may disagree over whether it is more meaningful to speak in terms of log odds or interpret results on the odds scale, or whether we should always present predicted probabilities (evaluated at the mean of covariates or using some other convention), but it has become less common to eschew any attempt at interpretation whatsoever. The simple process of grappling honestly with the question of what parameter values would seem noteworthy for a given model substantially improves the quality of subsequent analyses. Furthermore, should we discover that we have absolutely no idea how to interpret the basic components of our model to the extent that we can say—even as a purely subjective judgment—what parameter values would be substantively impressive to us, then perhaps we should ask ourselves whether we might not be better served by developing an alternate approach.

Whether one chooses to take the particular approach I have suggested here is far less consequential than a commitment to faithfully represent what matters from a theoretically informed, substantively oriented perspective. In fact, the simple approach that I have called the PASS-test is but a formalization of certain principles of sound statistical reasoning, which are no doubt second nature to many seasoned scientific professionals. If formally stated results of hypothesis tests continue to be the norm in scientific journals, it would restore some amount of balance to insist that scientific or practical significance be given at least equal attention in summaries of results such as those typically provided in tables and figures. Here I have outlined the bare essentials of what this sort of presentation might look like. Since presenting early versions of this work, I have encountered two sets of writings that address related themes within biostatistics and medical diagnostics literatures; these literatures, largely unknown among social scientists, offer a more extensive framework for just this sort of balanced approach. Two excellent books outline the process of generating and evaluating "informative hypotheses" (Hoijtink 2011; Hoijtink, Klugkist, and Boelen 2008) within a Bayesian perspective. Also especially relevant is work on *minimal clinically important difference*. Key articles include Jaeschke, Singer, and Guyatt (1989), Wells et al. (2001), Copay et al. (2007), and Revicki et al. (2008). The consequences of mistaking statistical significance for practical significance surely have higher stakes in such matters as pain reduction or medical risk assessment, so it is not surprising to see such areas begin to embrace this sort of approach. As political scientists, the immediate consequences of research may be less tangible, but if we believe that the fruits of our labor are nonetheless meaningful, we would do well to follow suit.

References

Altman, Morris, ed. 2004. "Statistical Significance" Special issue. *Journal of Socio-Economics* 33(5): 523–675.

Bakan, David. 1966. "The Test of Significance in Psychological Research." *Psychological Bulletin* 66(6): 423–37.

Berkson, Joseph. 1942. "Tests of Significance Considered as Evidence." *Journal of the American Statistical Association* 37(219): 325–35.

Carver, Ronald P. 1978. "The Case Against Statistical Significance Testing." *Harvard Educational Review* 48(3): 378–99.

Chow, Siu L. 1998. "Precis of Statistical Significance: Rationale, Validity, and Utility." *Behavioral and Brain Sciences* 21(2): 169–94.

Cohen, J. 1994. "The Earth Is Round ($p < .05$)." *American Psychologist* 49(12): 997–1003.

Copay, Anne G., Brian R. Subach, Steven D. Glassman, David W. Polly Jr., and Thomas C. Schuler. 2007. "Understanding the Minimum Clinically Important Difference: A Review of Concepts and Methods." *The Spine Journal* 7(5): 541–46.

DeGroot, Morris H., and Mark J. Schervish. 2002. *Probability and Statistics*. Boston, MA: Addison-Wesley.

Esarey, Justin, and Nathan Danneman. Forthcoming. "A Quantitative Method for Substantive Robustness Assessment." *Political Science Research and Methods*. http://jee3.web.rice.edu/riskstats.pdf.

Eysenck, Hans J. 1960. "The Concept of Statistical Significance and the Controversy about One-Tailed Tests." *Psychological Review* 67(4): 269–71.

Falk, Ruma, and Charles W. Greenbaum. 1995. "Significance Tests Die Hard: The Amazing Persistence of a Probabilistic Misconception." *Theory & Psychology* 5(1): 75–98.

Frick, Robert W. 1996. "The Appropriate Use of Null Hypothesis Testing." *Psychological Methods* 1(4): 379–90.

Gerber, Alan S., Donald P. Green, and David Nickerson. 2001. "Testing for Publication Bias in Political Science." *Political Analysis* 9(4): 385–92.

Gigerenzer, Gerd. 1993. "The Superego, the Ego, and the Id in Statistical Reasoning." In *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, ed. Gideon Keren and Charles Lewis. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., pp. 311–39.

Gigerenzer, Gerd. 1998. "We Need Statistical Thinking, Not Statistical Rituals." *Behavioral and Brain Sciences* 21(2): 199–200.

Gigerenzer, Gerd, and Zeno Swijtink. 1990. *The Empire of Chance: How Probability Changed Science and Everyday Life*. New York: Cambridge University Press.

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

Harlow, Lisa L., Stanley A. Mulaik, and James H. Steiger. 1997. *What If There Were No Significance Tests?* Mahwah, NJ: Erlbaum.

Hoijtink, Herbert. 2011. *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. Boca Raton, FL: CRC Press.

Hoijtink, Herbert, Irene Klugkist, and Paul A. Boelen. 2008. *Bayesian Evaluation of Informative Hypotheses.* New York: Springer.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2(8): e124.

Iyengar, Shanto, and Adam Simon. 1993. "News Coverage of the Gulf Crisis and Public Opinion." *Communication Research* 20(3): 365–83.

Jaeschke, Roman, Joel Singer, and Gordon H. Guyatt. 1989. "Measurement of Health Status: Ascertaining the Minimal Clinically Important Difference." *Controlled Clinical Trials* 10(4): 407–15.

Jeffreys, Sir Harold. 1961. *Theory of Probability*. Oxford: Oxford University Press.

Kish, Leslie. 1959. "Some Statistical Problems in Research Design." *American Sociological Review* 24(3):328–38.

Levin, Joel R. 1998. "What If There Were No More Bickering about Statistical Significance Tests?" *Research in the Schools* 5(2): 43–53.

Mayo, Deborah G., and Aris Spanos. 2006. "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction." *British Journal for the Philosophy of Science* 57(2): 323–57.

Meehl, Paul E. 1978. "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Consulting and Clinical Psychology* 46(4): 806–34.

Miller, Warren E., Donald R. Kinder, Steven J. Rosenstone, and the National Election Studies. 1999. *American National Election Studies, 1991 Pilot Election Study* [Data set]. Ann Arbor, MI: University of Michigan, Center for Political Studies.

Mogie, Michael. 2004. "In Support of Null Hypothesis Significance Testing." *Proceedings of the Royal Society of London: Series B, Biological Sciences* (Suppl. 3): S82–84.

Morrison, Denton E., and Ramon E. Henkel, eds. 1970. *The Significance Test Controversy: A Reader*. New Brunswick, NJ: Transaction.

Prior, Markus. 2009. "The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure." *Public Opinion Quarterly* 73(1): 130–43.

Rainey, Carlisle. Forthcoming. "Arguing for a Negligible Effect." *American Journal of Political Science.* http://onlinelibrary.wiley.com.libproxy.lib.unc.edu/doi/10.1111/ajps.12102/full.

Revicki, Dennis, Ron D. Hays, David Cella, and Jeff Sloan. 2008. "Recommended Methods for Determining Responsiveness and Minimally Important Differences for Patient-Reported Outcomes." *Journal of Clinical Epidemiology* 61(2): 102–09.

Rosenthal, Robert. 1979. "The 'File Drawer Problem' and Tolerance for Null Results." *Psychological Bulletin* 86(3): 638–41.

Rozeboom, William W. 1960. "The Fallacy of the Null-Hypothesis Significance Test." *Psychological Bulletin* 57(5): 416–28.

Schrodt, Philip A. 2006. "Beyond the Linear Frequentist Orthodoxy." *Political Analysis* 14(3): 335–39.

Serlin, Ronald C., and Daniel K. Lapsley. 1985. "Rationality in Psychological Research: The Good-Enough Principle." *American Psychologist* 40(1): 73–83.

Steiger, James H., and Rachel T. Fouladi. 1997. "Noncentrality Interval Estimation and the Evaluation of Statistical Models." In *What If There Were No Significance Tests?* ed. Lisa L. Harlow, Stanley A. Mulaik, and James H. Steiger. Mahwah, NJ: Erlbaum, pp. 221–57.

Thompson, Bruce. 2004. "The 'Significance' Crisis in Psychology and Education." *Journal of Socio-Economics* 33(5): 607–13.

Ward, Michael D., Brian D. Greenhill, and Kristin M. Bakke. 2010. "The Perils of Policy by p-value: Predicting Civil Conflicts." *Journal of Peace Research* 47(4): 363–75.

Wells, George, Dorcas Beaton, Beverly Shea, Maarten Boers, Lee Simon, Vibeke Strand, Peter Brooks, and Peter Tugwell. 2001. "Minimal Clinically Important Differences: Review of Methods." *Journal of Rheumatology* 28(2):406–12.

Wilkinson, Leland, and APA Board of Scientific Affairs Task Force on Statistical Inference. 1999. "Statistical Methods in Psychology Journals: Guidelines and Explanations." *American Psychologist* 54(8): 594–604.

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

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