Seven Deadly Sins of Contemporary Quantitative Political Analysis

Philip A. Schrodt

Pennsylvania State University

schrodt@psu.edu

Last update: 5/31/13

Word count: 9987, including bibliography

Acknowledgments: Paper originally prepared for the theme panel "A Sea Change in Political Methodology?" at the Annual Meeting of the American Political Science Association, Washington, 2 - 5 September 2010. This paper has benefitted from discussions with David Collier, Patrick Brandt, Hudson Meadwell, and John Freeman, as well as various conversations at subsequent presentations of the work. These individuals bear no responsibility for either the content or the presentation. Particularly the presentation. This research was supported in part by a grant from the U.S. National Science Foundation (SES-1004414). A more extended version of this paper, and related discussions, can be found at http://eventdata.psu.edu/7DS

Abstract

A combination of technological change, methodological drift and a certain degree of intellectual sloth, particularly with respect to philosophy of science, has allowed contemporary quantitative political analysis to accumulate a series of dysfunctional habits that have rendered much of contemporary research more or less meaningless. I identify these "seven deadly sins" as

- Garbage can models that ignore the effects of collinearity;
- Pre-scientific explanation in the absence of prediction;
- Excessive reanalysis of a small number of data sets;
- Using complex methods without understanding the underlying assumptions;
- Interpreting frequentist statistics as if they were Bayesian;
- A linear statistical monoculture which fails to consider alternative structures;
- Confusing statistical controls and experimental controls.

The answer to these problems is not to abandon quantitative approaches, but rather with solid, thoughtful, original work driven by an appreciation of both theory and data. The paper closes with suggestions for changes in current practice that might serve to ameliorate some of these problems.

The Problem

In recent years, I have found myself increasingly frustrated with the quantitative papers I am sent to review, whether by journals or as a conference discussant. The occasional unpolished gem appears, but the typical paper has some subset—often as not, the population—of the following irritating characteristics:

- A dozen or so correlated independent variables in a linear model;
- A new and massively complex statistical technique that is at best unnecessary for the problem at hand, as a simple t-test or ANOVA would suffice, and not infrequently completely inappropriate given the characteristics of the data and/or theory;
- Uses a data set that has been previously analyzed a thousand or more times;
- Is 35 ± 5 pages in length, despite producing results that could easily be conveyed in ten or fewer pages, as one finds in the natural sciences. That's for an R & R: First submissions are 60 ± 10 pages, with an apologetic note stating that the authors realize it may need to be cut slightly;

Not in the paper, but almost certainly under the surface, is a final factor

- The reported findings are the result of dozens—or more likely hundreds—of alternative formulations of the estimation.

Faced with such a paper, I do not believe the results. But realizing that the author[s] probably have children to feed, aging parents, small fluffy dogs, and will face a promotion-and-tenure committee that will simply count the number of refereed articles in their file, there is often little constructive I can say: This has become "normal science." "Change the topic, the data, the model, and the interpretation and maybe I'll find this interesting” while true, isn't all that useful. This sense is pervasive despite the fact that I am thoroughly positive and optimistic about the future prospects for quantitative political analysis. Just not the way it is being done now in the academic community.

In this deliberately polemical essay I will use the medieval trope of “seven deadly sins,” though I was hard-pressed to focus on only seven and my original list was closer to twenty. This work expands on points I made earlier in Schrodt (2006a), and also owes considerable intellectual debt to Achen (2002), who makes about two-thirds of the points I want to make here, albeit—as with King’s (1986) similar efforts—to little apparent effect. There will be a bias in this discussion—appropriate to the *Journal of Peace Research*—towards the fields in which I am most likely to review, the quantitative analysis of political conflict in both international and comparative forms.

Greed: Garbage can models and the problem of collinearity

Garbage can models are analyses where, in Achen’s [2002: 424] formulation, “long lists of independent variables from social psychology, sociology, or just casual empiricism, [are] tossed helter-skelter into canned linear regression packages.” Achen (2002) has reassuredly been *cited* 294 times according to Google Scholar (24 May 2013) and yet this remains the source of perhaps 80% of my distrust of contemporary quantitative research.

Achen's succinct “Rule of Three”— backed up with a number of methodological and technical justifications—asserts:

> With more than three independent variables, no one can do the careful data analysis to ensure that the model specification is accurate and that the assumptions fit as well as the researcher claims. … Truly justifying, with careful data analysis, a specification with three explanatory variables is usually appropriately demanding—neither too easy nor too hard—for any single paper. [Achen 2002: 446]

The specification of a linear model must always steer between the rock of collinearity and the hard place of omitted variable bias, the latter issue having been pursued in several recent expositions by Clarke (2005, 2012). Finding the right balance is challenging. As Daniel Kahneman (2011: chapter 21) discusses in considerable detail, the utility of simplicity echoes two decades of research going back to Dawes (1979) on “The robust beauty of improper linear models”—”improper” in the sense of “very simple.” This carries through the even older research of Meehl (1954) showing the superiority of simple statistical models, as well as the much older philosophical principal of “Occam’s Razor.” As with Achen [2002], Kahneman observes that neither Dawes (1979) nor Miehl (1954) has had the slightest impact on statistical practice in the social sciences.

Simple models have an edge for at least two reasons. First, complex models often “fit the error,” providing overly-optimistic assessments of the accuracy of the model for the existing data, but *decreasing* the accuracy of the model on any new data. Second, the nearly inevitable presence of collinearity in non-experimental social science models tends to increase the variance of the estimated coefficients as the number of independent variables increase. In contrast to the situation of controlled experimentation that motivated much of the development of modern statistical methods, where variables of interest can be varied independently, the political analyst typically confronts a situation where an assortment of equally plausible theories suggest several closely related (and therefore highly correlated) variables as possible causal factors.

This is compounded by operationalization issues. Economic concepts such as “price,” “interest rate,” or even “GDP” are unambiguously specified in a quantitative form even if measured with a substantial amount of error. In contrast, many important political science concepts—”power,” “legitimacy”, “authoritarianism,” or “civil war”—are qualitative and/or assessing a latent characteristic that has to be measured indirectly and can be operationalized in a variety of equally plausible ways.

Despite the availability of a number of well developed methods in psychology and testing which can estimate latent measures explicitly, and provide orthogonal (statistically independent) composite indicators no less, latent variable models are only rarely found in conflict research. Instead, analysts tend to simply throw an assortment of variables possibly relevant to the dependent variable into the model and hope that regression will magically sort it all out.

Linear models do not deal well with such situations. Collinearity may result in all of the relevant coefficients appearing to be individually insignificant or, quite frequently, will produce an estimate *opposite in sign* from the direct effect of the variable. Leave out a relevant variable—the aforementioned omitted variable bias problem—and its explanatory power is reflected in whatever related variables happen to be in the equation. Various diagnostics for this problem have been known for decades (Fox 1991) but typically one sees little more than a cross-correlation table and an assurance that none of the bivariate correlations are above 0.80, which is merely the point at which the bivariate correlation doubles the standard error.

In the absence of a strong linear effect in the main population, regression *amplifies* rather than isolates the influence of anomalous subpopulations, in the sense that the outlying subpopulation has a disproportionate effect on the values of the coefficients. How many published statistical results are actually the result of “hairball-and-lever” datasets consisting of a massive blob of uncorrelated cases with all of the significant coefficient estimates determined by a few clusters of outliers? We don't know, because very few published analyses check for this possibility and unlike the issue of collinearity, the number of possible subpopulations—Bell’s number—increases as a factorial, although Fox (1991) does provide diagnostics which could be used, and in most problems there are clear theoretical reasons to expect heterogeneous subpopulations.

In short, for many problems commonly encountered in political analysis, linear models aren't just bad, they are really, really bad. Arguably, it would be hard to design a worse set of potential side effects.

As a consequence, linear regression results are notoriously unstable—even minor changes in model specification can lead to coefficient estimates that bounce around like a box full of gerbils on methamphetamines. This is great for generating large numbers of statistical studies but not so great at ever coming to a conclusion. The orthodox response to this: “You have to resolve these inconsistencies on the basis of theory.” But usually the whole point of doing the test was to empirically differentiate competing and equally plausible theories! The cure becomes equivalent to the disease, a problem we will further explore in the incompatibilities between the hypothetical-deductive method and the frequentist statistical paradigm within which these linear models are embedded.

Pride: Pre-scientific explanation in the absence of prediction

One of the most mystifying—and exasperating and self-indulgent—tendencies in the quantitative international relations (IR) community over the past two or three decades has been the disparaging of prediction as the criteria for validating a model, instead preferring “explanation,” incongrously if conveniently defined as coefficient estimates barely distinct from zero as estimated in profoundly problematic linear models on a null hypothesis the researcher has no reason whatsoever to believe is true. Papers and articles that attempt to forecast are simply dismissed by the discussant/referee with a brusque “That’s only a forecast.”

This is a perfectly understandable human impulse—if you can't make the goal with frequentist models (Ward et al 2010), move the goal posts—though less understandable in this context since it is quite straightforward to develop successful predictive models of political conflict behavior, the Political Instability Task Force (PITF; Goldstone et al., 2010) and the Integrated Conflict Early Warning System (ICEWS; O’Brien 2010) being conspicuous recent examples. Furthermore, this is certainly not where the quantitative IR community started: The early proponents were motivated to develop models that were predictively accurate in hoped that such knowledge would reduce the probability of their day being ruined by a US-Soviet thermonuclear conflagration. There has been a nearly continuous interest in prediction going back to the early 1970s and continuing to the present (McClelland 1969, Choucri and Robinson 1979, Vincent 1980, Hopple et al. 1984, Esty et al. 1998, Davies and Gurr 1998, Pevehouse and Goldstein 1999, Schrodt and Gerner 2000, King and Zeng 2001, Bueno de Mesquita 2002, Schrodt 2006b, Schneider et al. 2010, Weidmann and Ward 2010, Brandt et al. 2011).

This philosophical position has puzzled me from the first time I encountered it. In the natural sciences, successful forecasts are the epitome of validation of a theory, and some successful predictions—for example Edmond Halley's forecast of the return of the eponymous comet, or Sir Arthur Eddington’s 1919 confirmation of Einstein’s prediction of the deviation of starlight during a total eclipse—are considered landmarks in the history of science. In the social sciences, one finds industrialized countries spending hundreds of millions of dollars on data collection and econometric modeling in order to provide economic forecasting. The accuracy (and influence) of opinion polls is now sufficiently high that their publication in the days prior to an election is now regulated in many democracies, and *New York Times* analyst Nate Silver famously predicted every single electoral vote in the 2012 US presidential election in an environment where many high-profile qualitative pundits were predicting a landslide victory for Republican Mitt Romney.

I have been asking proponents of this position, for years, to provide a source for it, to no avail. So in the absence of a specific argument to refute, one can only provide evidence to the contrary. Even this is difficult as there is a complete disconnect between opposition to prediction and 20th century philosophy of science which, for the most part, was coming out of the deterministic predictive traditions of Newton, Laplace, Maxwell and took for granted the centrality of prediction in scientific practice.[1]

Nonetheless, with a bit of digging one can find some succinct arguments; these are covered in much more detail in Schrodt (2010). Carl Hempel's classic covering law essay is titled "Explanation and Prediction by Covering Laws" (Hempel 2001) suggesting that for Hempel the two go together in a properly scientific theory, and throughout that essay Hempel unambiguously treats explanation and prediction as equivalent. The logical positivists, being rather rigorous

[1] Quantum mechanics introduced randomness at the sub-atomic level but, Orme-Johnson et al (1988) notwithstanding, the deterministic laws of the eighteenth and nineteenth centuries remained valid at the level of directly observable phenomena.

logicians (sometimes maddeningly so), would of course be completely appalled at the notion that two things could simultaneously be equivalent and one of them weaker than the other.

Hempel and Oppenheim (1948) put this into a more complete context:

> It may be said, therefore that an explanation of a particular event is not fully adequate unless its *explanans*, if taken account of in time, could have served as a basis for predicting the event in question. Consequently, whatever will be said in this article concerning the logical characteristics of explanation or prediction will be applicable to either, even if only one of them should be mentioned.
>
> Many explanations which are customarily offered, especially in *pre-scientific discourse* [emphasis added], lack this potential predictive force, however. Thus, we may be told that a car turned over on the road "because" one of its tires blew out. Clearly, on the basis of just this information, the accident could not have been predicted, for the *explanans* provides no explicit general laws by means of which the prediction might be effected." (Hempel and Oppenheim (1948: 138-139)

The critical insight from Hempel (and the logical positivists more generally: see Quine 1951) is that explanation in the absence of prediction is not scientifically superior to predictive analysis, it isn't scientific at all! It is, instead, "pre-scientific."

Yet this argument has had little impact: I have been persistently challenged to provide "a source more recent than Hempel" to support this contention (while, I would repeat, the significance-test-based "explanation" camp does not feel an obligation to provide any philosophical support whatsoever for their position). I will, at this point, admit failure. Putting me in the same position as those who cannot find any support more recent than the mid-seventeenth century for the heliocentric model of the solar system.

The pre-scientific character of explanation in the absence of prediction can be illustrated by considering the phenomenon of lightning. For many centuries, the well-accepted and quite elaborate explanation among some Northern European cultures was that lightning bolts were hurled by the Norse god Thor. For believers in Thor, this “explanation” had all of the intellectual complexity and coherence of, say, rational choice or balance of power theory, and certainly more entertainment value. And it had some useful predictive value—Thor, it seems, liked to use isolated trees and mountain peaks for target practice, and it was best of avoid such places when lightning was about.

Yet the “Thor theory of lightning” failed some critical tests, notably when the Anglo-Saxon missionary St. Boniface chopped down the sacred Thor's Oak in Fritzlar (modern Germany) in 723 and Thor failed to come to the oak's defense. More generally, knowing the ways of lightning required knowing the mind of Thor (much as rational choice and balance of power theory requires knowing the unknowable utilities of political actors), and was of limited practical utility.

Contrast this with the scientific understanding of lightning that developed in the mid-18th century, through the [distinctly hazardous] experiments of Franklin in North America and Dalibard and De Lors in France. Both established that lightning was a form of electricity. Deductively, if lightning is electricity, it will flow through good electrical conductors such as iron and copper better than through poor conductors such as wood and stone. Hence metal lightning rods could protect buildings from lightning, a practical and empirically verified prediction. Sven sacrifices goat to Thor; Sven's barn burns down. Helga installs lightning rod; Helga's barn survives. Electricity theory good; Thor theory not so good.

There is, of course, a place for pre-scientific reasoning. Astrology provided part of the empirical foundation for astronomy, and no less a scientific mind than Newton devoted a great deal of attention to alchemy. This comparison of purely "explanatory" theories to astrology is anything but a cheap shot: astrology has virtually all of the components of a legitimate scientific enterprise *except* predictive validity, and the challenge of differentiating astrology from orthodox science has been an issue in philosophy of science since the time of Francis Bacon. Furthermore pre-scientific heuristics of, say, rational choice theory may provide some insights, much as chatting with people in war zones, or immersing oneself in dusty archives can provide insights. But none of these are scientific: Only predictive models are scientific.

Sloth: "Insanity is doing the same thing over and over again but expecting different results."[2]

The genius of the scientific method is that it can allow for incremental advances to be made using relatively routinized procedures implemented systematically by a large number of people. But this also carries a risk: Progress comes to a halt when those routine increments to knowledge have been exhausted.

We are presently in a situation of limited progress, at least in the refereed journals: Most of the easy things appear to have been done, and the routinized procedures only contribute to further confusion. Too many findings can be undone by a slightly different analysis of the same data, and

[2] Usually attributed to Albert Einstein, and occasionally Benjamin Franklin; in fact it is apparently due to one rather contemporary Rita Mae Brown in 1983.

even experts—to say nothing of the general public—have a difficult time deciding between them. I believe very little of what I'm reading in the journals, and this is not a good thing.

There is an old saying in the natural sciences that you should try to write either the first article on a topic or the last article. Rummel's empirical work on the democratic peace was interesting (Rummel 1979, and much more exhaustively, http://www.hawaii.edu/powerkills/), as were Russett's (1993) effort to bring the hypothesis into the academic—and later, policy—mainstream. The hypothesis certainly needed empirical testing, and the data set by Oneal and Russett (1999; Russett and Oneal 2001) became the canonical mode for doing this. Quite possibly, Oneal and Russett missed something really important and a few additional articles using their data set would be worthwhile. But 161 articles?: This is the *Web of Science* count of the citations to that article on 24 May 2013, up from 113 since I first presented this argument in August 2010. Most of those articles are just minor specification, operationalization or methodological variations on the original, collinearity-fraught data set, so all we are seeing are essentially random fluctuations in the coefficient values and standard errors.

Not all of those citations, of course, involve a reanalysis of the data. Let's assume, conservatively, that only 50% involve re-analysis. Let's also assume—this may or may not be accurate—a 1-to-3 yield rate of research papers to publications, and finally—this is could easily be underestimated by a factor of five—the average paper resulted from twenty analyses of the data using various specifications. This means—with very serious consequences to frequentist practice—that the data have been re-analyzed about 4,800 times.

Data cannot be annealed like the steel of a samurai sword, becoming ever stronger through each progressive application of folding and hammering. Folding and hammering data only increases the level of confusion. There is only a finite amount of information in any data set and when competently done, the first five or ten articles will figure out those effects. I defy the reader to provide me with a single example where the *reanalysis* of a data set after it had been available for, say, two years, has produced robust and important new insights except under circumstances where the assumptions underlying the original analysis were flawed (e.g. not taking into account time-series, nesting or cross-sectional effects).

Nor, except in very unusual circumstances—usually when the original indicators contained serious non-random measurement errors and missing values—will adding or substituting a closely related indicator to the earlier data set make any consistent difference. Methods robust to the existence of collinearity such as cluster analysis and principal components will just ignore this as they've already detected the relevant latent dimensions in the existing indicators. Brittle methods—regression and logistic—will go berserk and rearrange the coefficients based on subtle interactions (sometimes, literally, round-off error) occurring during the inversion of the covariance matrix. None of this has any meaningful relation to the real world.

The most tragic aspect of this process is the opportunity cost in terms of the data sets that are insufficiently analyzed. Systematic data collection is something we *really* know how to do now: The number of well-documented and reasonably thorough data sets now available and relevant to the study of political behavior is astonishing, a completely different situation than we had thirty years ago. An APSA-sponsored conference at Berkeley in Fall 2009 on the cross-national quality of governance identified some 45 data sets potentially relevant to the issue; a compendium of open-source indicators available to PITF identified almost 3,000 variables.

Furthermore, data collection is *not* a monoculture: we are now in a situation where we can systematically evaluate the extent to which we get similar results from multiple convergent indicators. But instead we largely see the reanalysis of a small number of canonical data sets, even when those have well-known problems (e.g. the intermediate categories in the democracy-autocracy measures in *Polity*).

Lust: Using complex methods without understanding the underlying assumptions

For a time during the 1990s, there was a great deal of concern expressed about the possibility of “unit roots”—essentially random walks—in political science time series, and virtually every article I was sent (or tried to publish) using event data saw some reviewer asking whether the series had been checked for unit roots.

There were two problems with this. First, while the unit root hypothesis was theoretically credible in the fields of econometrics where it was originally developed, there were absolutely no reasons to expect unit roots in event data, and furthermore, the bounded character of news reporting at the time made this virtually impossible. (This applied even more dramatically in the domain of aggregate public opinion, where authors also had to endure the demand for unit root tests.) Second, and more problematic, the commonly used tests for unit roots had extremely low power—the probability of correctly rejecting the null hypothesis when it is false—when a series was highly autocorrelated, but still not a random walk. Exactly the situation found in many event data (and public opinion) series. Researchers were nonetheless constantly asked to perform these atheoretical and very flawed tests because, well, the econometricians were doing it on stock market data. Like any other mania, this obsession eventually burned itself out, but not before delaying progress in the field a few years.

There is nothing unique about unit root tests in this regard: over the years, I have seen countless examples where a paper uses a complex method—usually developed in a field distant from political science, and usually with little evidence that the researcher actually understands the method—and applies it in situations which clearly violate the assumptions of the method. Often as not, “everyone is doing it” carries the day with the journal editors: methodologists are such boring nags, worrying about fundamentals and all that other useless stuff. One will then see a cascade of equally bad papers until someone notices that most of the resulting estimates are incoherent and might as well have been produced by a random number generator, and we quietly sidle off to the next set of mistakes.

Once again, Achen (2002) has [fruitlessly] covered this ground rather thoroughly. I'm just saying it again.

With just a couple more observations.

Complex models are not always inappropriate, and in some situations they are clearly superior to the simpler models they are displacing. One of the most conspicuous examples of this would be the use of sophisticated binary time-series cross-sectional estimation in IR following the publication of Beck et al (1998). Quantitative IR was analyzing a lot of such data; existing methods could easily incorrectly estimate the standard errors by a factor of two or more. The revised methodology, while complex, was completely consistent with the theory and data, and consequently its use was wholly appropriate. The increase over the past two decades in the use of hierarchical linear models in situations of nested observations would be another good example. The sophisticated use of matching methods probably also qualifies, as does imputation when it is consistent with the data generating process.

However, for each of these success stories, there are numerous cases where one sees complexity for the sake of complexity, in the hopes (often, alas, realized) that using the latest technique (conveniently a few mouse-clicks away on CRAN) will get your otherwise rather mundane analysis over the bar and into one of the five [sic] Sacred Top Three Journals and in help to get you tenure. But, in fact, the complex technique probably makes at best marginal changes to your coefficient estimates and standard errors because it is only effective if you can correctly specify things you probably don't know such as the variance-covariance matrix of the errors, or the true propensity function in a matching problem.

In the meantime, this bias towards complexity-for-the-sake-of-complexity (and tenure) has driven out more robust methods. If you can make a point with a simple difference-of-means test, I'm more likely to believe your results because the t-test is robust and requires few ancillary assumptions (with the key one is usually provided by the Central Limit Theorem). Running a regression with only dummy independent variables? (yes, I've seen this…): What you really want—actually, what you've already got—is an ANOVA model (very robust, though rarely taught in political science methodology courses). You have a relatively short time series and good theoretical reasons to believe that both the dependent variable and the error terms are autocorrelated (and in most political behavior, this will be the case)?: You can worship at the shrine of Box, Jenkins and Tiao and wrap your variables into transformational knots that even a massage therapist couldn't unwind, or you can just run OLS, but either way, you aren't going to be able to differentiate those two effects. But with a simple model at least you will be able to interpret the OLS coefficients.

Upshot: use the simplest statistical method that is consistent with your theory and data. Rather as kindly Dr. Achen suggested more politely a decade ago.

Wrath: If the data are talking to you, you are a Bayesian

At the pedagogical and mainstream journal level in political science we have legitimated a set of rather idiosyncratic and counterproductive frequentist statistical methodologies. These are the hoary legacy of an uneasy compromise that came together, following bitter but now largely forgotten philosophical debates by Fisher, Neyman, Pearson, Savage, Wald and others in the first half of the 20th century (Gill 1999, McGrayne 2011), to solve problems quite distant from those encountered by most political scientists. As Gill points out, this Fisher-Neyman-Pearson "ABBA" synthesis—"Anything But Bayesian Analysis"—is not even logically consistent, suggesting that one of the reasons our students have so much difficulty making sense of significance tests is that in fact the tests don't make sense.

The pathologies resulting from frequentism applied outside the rarified domain in which it was originally developed—induction from random samples—are legion and constitute a sizable body of statistical literature: Freedman (2005) and Freedman et al (2009) is as good as place as any to start; Ziliak and McCloskey (2008) provide additional critiques. To call attention to only the most frequent [sic] of these problems as they are encountered in political science:

1. Researchers find it nearly impossible to adhere to the correct interpretation of the significance test. The p-value tells you only the likelihood that you would get a result under the [usually] completely unrealistic conditions of the null hypothesis. In fact, outside of a purely frequentist mindset, one usually wants to know the magnitude of the effect of an independent variable, given the data. That's a Bayesian question, resolved with the posterior distribution of the coefficient. Instead we see—constantly—the p-value interpreted as if it gave the strength of

association in the ubiquitous Mystical Cult of the Stars and P-Values which permeates our journals.

The frequentist paradigm—leave aside the internal contradictions with which we have somehow coped for close to a century—applies fairly well in the two circumstances for which it was originally developed: random samples and true experiments. These are encountered in *some* important areas of political science research, survey research being the most obvious. But there are large swaths of political science where they do not apply, notably pretty much the whole of IR. In these situations, usually one is studying a population rather than a sample, and while one can go through no end of six-impossible-things-before-breakfast gyrations—measurement error, alternative universes, etc.—to try to justify the use of sample-based methods on populations, they are fundamentally different. This debate has a very long history: see Morrison and Henkel (1970).

2. The ease of exploratory statistical computation has rendered the traditional frequentist significance test all but meaningless. Alternative models can now be tested with a few clicks of a mouse and a micro-second of computation (or, for the clever, thousands of models can be assessed with a few lines of programming). Virtually all published research now reports only the final tip of an iceberg of dozens if not hundreds of unpublished alternative formulations. In principle significance levels could be adjusted to account for this; in practice they are not, and the sheer information management requirements of adjusting for the 4,800+ models run in multiple research projects on the Oneal-Russett data (or ANES, or Polity, or GSS, or EuroBarometer, or the Correlates of War instantiated in EuGENE) render such an adjustment impossible.

3. Finally—for this list—there is a very serious inconsistency between the frequentist presuppositions and hypothetical-deductive, theory-driven analysis ("micro-foundations" in Achen's terminology). Nothing wrong with theory: theory is what keeps *parakeets_per_capita* out

of our models. Well, most models. But if your model is theory-driven, the rejection of the null hypothesis doesn't tell you anything you didn't know already—your theory, after all, says that you expect the variable to have at least *some* effect, or it wouldn't be in the model in the first place. Rejection of the null hypothesis merely confirms this.

If one were operating in a strict falsification framework acceptance of the null hypothesis might be useful. Though only if somehow one could get around measurement, specification and collinearity problems, the low power of the significance test in distinguishing between multiple closely related specifications, and actually believe the results of a single test rather than trusting your intuition and estimating yet another alternative formulation of the model when the coefficient estimates seem just too weird (that's Bayesian again!). Still, if in numerous alternative formulations a variable still isn't significant, that is probably fairly good evidence to conclude it is not relevant—unless it is one of those *Night of the Living Dead* zombie hypotheses like the diversionary theory of war—so such tests can provide occasional progress.

But as a long literature has established—this was one of the jumping-off points for Kuhn (1962)—scientific inquiry, while accepting the *principle* of falsification, only rarely proceeds using strict falsification norms. Instead, the tendency is to do extensive exploratory work and substitute paradigms only when a superior alternative is firmly established. In the stochastic realm of social behavior, the failure to reject a null hypothesis in a single instance—nominally how the frequentist approach works—tells us almost nothing.

The alternative, of course, is Bayesian approaches. At some levels these are already widely accepted: one will find in most statistics departments at least half of the researchers are Bayesian. Bayesian approaches are common in the consistently top-ranked—by impact factor—*Political Analysis* (for example Hoeting et al 1999, Lock and Gelman 2010, Montgomery and Nyhan 2010, Alvarez et al 2011, Grimmer 2011, Montgomery et al 2012) but not in more mass-market venues such as the *American Political Science Review* and *American Journal of Political Science*, which are overwhelmingly frequentist, nor in most quantitative IR work.

The Bayesian alternative solves numerous problems: it is logically coherent, and as such it can provide the basis for a proper theory of inquiry, it cleanly solves the issue of the integration of theory and data, it is agnostic on the issue of populations versus samples, and it provides a straightforward, if still underutilized, method of integrating informal *a priori* information with systematic data-based studies. Bayesian approaches corresponds to how most people actually think, no small advantage when developing models of human behavior.

The downside to Bayesian approaches is their mathematical and computational complexity. The latter now has purely technological fixes, though the prospect of substituting 48-hour WinBUGS runs for OLS is less than appealing. Furthermore, while talking the Bayesian talk, the quantitative community is still generally not walking the walk through the use of informative priors. Do we need strict Bayesianism, or merely a less restrictive "folk Bayesianism" (McKeown 1999) that drops the most objectionable aspects of frequentism but still allow some pragmatic lessons-learned from the past century of statistical work? This is very much an on-going debate in the statistics community—Andrew Gelman's blog http://andrewgelman.com/ is an excellent place to follow it—and we should be part of that debate, not looking away from it. Inside every confused graduate student or assistant professor questioning why it makes any sense to compute a significance test on a population (hint: it doesn't…), there is a Bayesian struggling to break free.

Gluttony: Enough already with the linear models!

Even the most cursory glance at quantitative studies in the mainstream journals over the past twenty years will show that we have become a statistical monoculture: virtually all analyses are done with variations on linear regression and logit.

Linear models are a perfectly good place to start: They are computationally efficient, well understood, the estimators have nice asymptotic properties, and, using a Taylor expansion, the linear functional form is a decent first approximation to pretty much anything. Anyone teaching quantitative methods will have a folder of scattergrams showing real-world examples that plot out nicely along a line, perhaps with a few interesting and plausible outliers.

But monocultures always have the same unhappy ending: parasitism, disease and eventual collapse. Parasitism in this context is the individual, *homo significantus,* who, year after year, grinds out articles by downloading a data set, knocks out a paper or two over the weekend by running a variety of specifications until—as will invariably occur—some modestly interesting set of significant coefficients is found, and through a network of like-minded reviewers and the wearing down of journal editors, publishes the results. Dear reader, do we not all know at least one person fitting this description?

The problems with this monoculture have been detailed elsewhere in this essay; the point is that there are alternatives. Consistent with my monoculture metaphor, social science statistical work was far more methodologically rich, creative and likely to adjust tests—grounded in probability theory—to specific theories, problems, and data in the past than it is now (see for example Anderson 1958, Lazarfeld 1937, Richardson 1960). Arguably, we are also lagging well behind the non-academic data analysis sector: see *Economist* 2010, *Science* 2011, Schrodt 2009 and the work of both PITF and ICEWS. Just like the poor city kid who has never seen a tomato that is not a pasty yellow-pink and the consistency of a croquet ball, too many political scientists think "statistics" equals "regression" and as a consequence believe, for example, that inference is impossible if the number of potential explanatory variables exceeds the number of cases. In fact almost all human inference occurs in such situations; this is only a limitation in a world of linear models.

The number of methods we are *not* using is stunning. Correspondence analysis (CA) is a method almost unseen in North American research, but is every bit as much a sophisticated data reduction method as regression, can be derived from a variety of assumptions, and is available in a myriad of variations. Support vector machines (SVM) provide another example. These are the workhorse of modern classification analysis, well-understood, highly robust, readily available, and yet generally absent in political analysis except in applications to natural language processing.

This is the tip of the iceberg. Just sampling from three current texts on computational pattern recognition—Duda et al. (2001), Bishop (2006), and Theodoridis and Koutroumbas (2009)—one finds addition to the methods discussed above multiple variations on

- neural networks
- Fourier analysis
- principal components
- hidden Markov models
- sequential, functional, topological and hierarchical clustering algorithms
- latent variable models
- genetic algorithms and simulated annealing methods

I am not advocating these alternative methods as novelty-for-the-sake-of-novelty—that would be as dysfunctional as the complexity-for-the-sake-of-complexity. But at a minimum these techniques provide alternative structures for determining regularities in data—just because many things are linear doesn't mean that *everything* is linear—and in many cases, they are better suited than linear methods in dealing with issues commonly found in political science data.

For example, a number of these methods are completely workable in situations where the number of independent variables is greater than the number of cases, and most clustering algorithms are ambivalent as to whether variables are correlated. Many of these methods can use missing values as a potential classifier, which is very relevant in situations where data fail the missing-at-random test (for cross-national data, almost all situations).

A consistent criticism I've received is that this advice contradicts the point made in the "Lust" section that one should not seek out complex models. This confuses the issue of complexity—in the sense of the underlying assumptions of the model—with the issue of whether a method is commonly taught and consequently understood in the field. By almost any measure, SVM and decision-tree methods are *simpler* than many of the regression based methods one commonly encounters in contemporary work, and even the more complicated methods are of comparable complexity. Furthermore, these methods are increasingly used in applied work—PITF routinely tests a variety of models from the frequentist, Bayesian and machine learning fields when developing predictive models—though not, for the most part, in the academic journals in political science. All of these methods are readily available, for only the cost of the time spent learning them, in *R,* as well as user-friendly packages such as Weka (http://www.cs.waikato.ac.nz/ml/weka/).

Envy: Confusing statistical controls and experimental controls

One of the more interesting exercises in my career was a methodological rebuttal (Schrodt 1990; see also Markovsky and Fales 1997) to an analysis published in the *Journal of Conflict Resolution* that purported to establish the efficacy of Transcendental Meditation, at a distance, in reducing levels of political violence (Orme-Johnson et al. 1988). While I found multiple issues with the analysis (as did Markovsky and Fales), the key element—in this and other TM studies—was their interpretation of the inclusion of additional independent variables as "controls."

Orme-Johnson et al were hardly out of line with prevailing practice to do this: such characterizations are all too common. But except in carefully randomized samples—and certainly not in populations—and with sets of statistically independent variables (which in the social science research, outside of experimental settings, almost never exist) statistical "controls" merely serve to juggle the explained variance across often essentially random changes in the estimated parameter values. They are in no way equivalent to an *experimental* control. Yet too frequently these "control variables" are thrown willy-nilly into an estimation with a sense that they are at worst harmless, and at best will prevent erroneous inferences. Nothing could be further from the truth.

This is another situation where we have gradually, and without proper questioning, drifted into an mode of expression which while comfortable—randomized experiments are the gold standard for causal inference—is simply dead wrong in the contexts where we apply it: the estimation of linear coefficients from sets of correlated independent variables measured across inhomogeneous populations.

For a number of years, the first exercise in my advanced multivariate methods class (you don't want to do this in the introductory class) was to give the students a cross-national data set and have them find the most ludicrous model possible in terms of obtaining significant coefficients on nonsensical independent variables due to spurious correlation or, more commonly, collinearity and outlier effects. No student had the slightest problem doing this. None, to my knowledge, tried to publish any of these models, but I sense that our journals are effectively filled with similar, if inadvertent, exercises.

The typical model presented in quantitative conflict analysis involves three or four primary explanatory variables, still often presented in the archaic H_1, H_2, H_3… form accompanied by eight to ten additional variables designated as "controls." These "controls" actually have little or nothing to do with classical experimental controls, and are in fact much closer to the "ancillary assumptions" which doomed the logical positivist effort at bringing closure to the scientific enterprise.

Calling these "controls" doesn't change how the estimation software treats them: the estimation routine is utterly indifferent as to whether you call the variables *explanatory, control, Tinkerbell* or *menneskerettighetsorganisasjonssekretæren.* Nor does the software care that you put the "explanatory variables" in the first lines of your regression table and the "controls" beneath them.

As the proud parents of H_1, we envision it at the front of the stage, singing its little heart out. But in point of fact, it's back in the corner, standing on tiptoes, saying in a high squeaky voice "Look at me, look at me, PLEASE look at me!" But the controls are often the big guys—in the conflict literature, tough bruisers like contiguity, GDP/capita, and conflict_{t-1}, and...well, those big guys will probably just steal poor little H_1's lunch money, and H_1 will not have a nice day. In an ideal world, we can see this going on, but in *our* world, where "controls" are likely as not collinear, pretty much anything can happen. Life as just another variable inside the X matrix is tough, and $(X'X)^{-1}X'y$ is cold and heartless.

The derived wisdom on the list of acceptable “controls”—generally a small subset from a very much larger universe of theoretically plausible and imminently measurable variables—is largely determined by prior practice and data availability. In a contemporary analysis, these variables will be presented in anywhere from a half dozen to two dozen variations over the course of a paper. Those norms have proven quite robust in predicting the content of conference presentations, job talks, and articles, particularly at the pre-publication stage. I have found it necessary to instruct my grad students not to start giggling upon seeing some poor misguided job candidate proudly display an unintelligible table perfectly matching these parameters.

The other side of this coin—and yet another pathology of frequentism—is the assumption that statistical significance has causal implications. Fortunately, our understanding of this is considerably more sophisticated than it was two decades ago—as expressed, for example, in the causal inference focus of the 2009 Society for Political Methodology summer meeting at Yale—but the error still permeates discussions in the discipline. In a suitably controlled and randomized experiment, a strong variable effect will usually (leaving aside the possibility of spurious correlation due to omitted variables) translate into a predictable effect on the dependent variable. This is not true in an equation estimated on noisy data from a population.

This has serious implications. Much of the early PITF work (Esty et al 1998) proved to be a dead-end because the variables which were statistically significant did not translate into any gains in prediction, a problem that has plagued the quantitative analysis of causes of political conflict more generally (Ward et al 2010). Only when PITF methodology shifted to modes of assessment that specifically measured predictive validity—for example split-sample testing and classification matrices—were the models able to transcend this problem. ICEWS, presumably learning from the experience of PITF, used predictive evaluation as the criteria from the beginning.

What is to be done?

Despite this long list of criticisms of current practice, I should adamantly assert that I'm not suggesting throwing out the scientific method and reverting to a fuzzy-wuzzy “I’ll know it when I see it (well, maybe...whatever...)” approach or, worse, to a postmodern narcissistic nihilism that denies the possibility of an objective reality. Given the number of well-studied pathologies in human intuitive reasoning (Vertzberger 1990, Tetlock 2005), even among experts, we need systematic methods to figure out political behavior. Instead, I suggest that we take these and earlier criticisms as the guideposts towards the development of a new and more sophisticated philosophy of inference specifically designed for political analysis, rather than simply adopting whatever worked in the quality control department of the Guinness Brewery in 1908.

It has taken a while to get ourselves collectively lost in this dismal swamp, and it will take a while to get out, but let me suggest four points where we should focus.

1. There should be zero tolerance—among discussants, reviewers, editors, department heads, search committees and tenure committees—for bad practices that we’ve always known we shouldn’t be doing. Garbage can models are meaningless; significance is not the same as causality; don't use methods that are inappropriate for your theory and data. In a couple instances—the self-satisfied drift into pre-scientific “explanation” at the expense of prediction, and the tolerance of nearly infinite re-analysis of a small number of data sets—there’s probably a serious need to go back and clean up the collective mess, and some of that is in a larger professional context (e.g. lazy tenure committees who simply count publications).

2. Begin the transition away from frequentism into more recent methods which actually do what we think they do. Starting with shifting to Bayesian methods—at the very least adopting folk Bayesianism but also some of the more user-friendly technical methods such as Bayesian model averaging (Montgomery and Nyhan 2010, Montgomery et al 2012; BMA also makes quick work of garbage-can models). Reserve frequentism for those cases, which are rare in conflict studies, where data are from a random sample and there are theoretical reasons to believe a coefficient might be zero. Use contemporary case-control and matching methods (Sekhon 2008, Hainmuller 2012, Iacus et al 2012) —which clearly distinguish between control and explanatory variables—rather than dumping everything into an undifferentiated and collinear regression matrix.

Meanwhile, update the core graduate methods curriculum, which has not changed significantly since the 1960s. One need not apologize to graduate students about teaching confidence intervals, or distributions, or the Central Limit Theorem, or probability theory. . . again, don't throw out the proverbial baby with the proverbial bathwater. But the frequentist approach as a whole does not make logical sense, particularly, as is common in IR, when we are dealing with populations rather than samples. Nor is it possible to reconcile a preference for the deductive-hypothetical method with the frequentist null hypothesis approach: If we have theory guiding our models, then the *tabula rasa* of the null hypothesis is both intellectually dishonest—we are claiming to start from a mindset which we most certainly do not have—and the information it provides us is generally useless.

3. Open the journals to alternatives to linear regression and logit, both to models that are simpler and those which are more novel: We need to seek a middle way between post modernist ascetic nihilism and technical virtuosity solely for the sale of novelty, while fending off Mara's hordes of anonymous taunting reviewers and the weekend wonders who equate analysis with

mouse clicking. This, however, requires *more* sophisticated training in methods, not the dumbing down of the quantitative curriculum proposed by Mearsheimer and Walt (2013), who wish instead to see the revival of pompous and long-discredited "grand theories" expounded by the likes of, well, Mearsheimer and Walt.

4. Either provide a philosophical justification for the primacy of significance-test-based "explanation"—a *very* steep hill to climb—or join the rest of the sciences and the policy community and return to an emphasis on prediction, with the appropriate adjustment in methods. I emphasize a *philosophical* justification, not merely a lame Kuhnian/social constructivist "This is how we've been doing things, therefore it is normal science, therefore it is science." So was astrology.

To conclude: Since 2010, I've presented these ideas at a variety of venues, and the response of the audience is predictable. People under the age of 35 love it. People over the age of 45 hate it. People between the ages of 35 and 45 are ambivalent: "Well, you're right, but what is this going to do to my research?" As a harbinger of the future of our discipline, that response leaves me guardedly optimistic.

But only guardedly. The institutional inertia of the entrenched academic interests is so pervasive that we could also be entering a phase where scientific innovation occurs, for the most part, outside of academia. The institutional response to Achen (2003) was, of course, the notorious journal-length methodological suicide note, *Conflict Management and Peace Science* Vol. 22, No. 4, which, like J.R.R. Tolkien's Gollum leaving the sunlit world for a lonely life in subterranean darkness, reads like "Precious, oh precious garbage can models. Evil Achens wants to take away garbage can models...no, no, we won't lets them...precious garbage can models…" The peer-review system—as recently described to me by an editorial assistant at a major IR journal—has degenerated into a stultifying version of *The Hunger Games*, a domain where assistant professors, armed only with anonymous reviews, strive to eliminate each other in the vicious pursuit of an ever-diminishing supply of tenured positions while their elders, secure in such positions, watch in amusement. The fossilized detritus of the review process is released, after a two-year delay, into a world that has moved on at internet speeds. A student trained in the standard academic methods curriculum would be completely lost in the world of Bayesian predictive models of PITF or ICEWS, or much of the geospatial work of the Peace Research Institute Oslo or Amnesty International. The center of innovation has shifted.

We've seen this happen before: The university system opted out of the scientific methods pioneered by Bacon and Descartes in the early seventeenth century, retaining a static late-medieval curriculum until the diffusion of the Humbolt reforms in the nineteenth century. An intellectual lag lasting a mere two to three centuries.

With contemporary networked communications, I believe change will happen more quickly this time. But for me, not quickly enough. The unexpected consequence of writing Schrodt (2010) was finding myself no longer capable of (or credible) teaching students to run garbage can models but still realizing that if they wanted secure careers, they had to run garbage can models. Lots of garbage can models. Rather than live with this contradiction, I've resigned from the gilded cage of my tenured academic position, and will henceforth make my way developing quantitative models in the more open world of policy-oriented forecasting. It's a magical world out there: let's go exploring! [http://bestofcalvinandhobbes.com/2012/02/final-calvin-and-hobbes-its-a-magical-world-lets-go-exploring/]

References

Christopher Achen. 2002. Toward a new political methodology: Microfoundations and ART. *Annual Review of Political Science* 5: 423-450.

R. Michael Alvarez, Delia Bailey, and Jonathan N. Katz. 2011. An Empirical Bayes Approach to Estimating Ordinal Treatment Effects. *Political Analysis* 19,1: 20-31

T. W. Anderson. 1958.*The Statistical Analysis of Time-Series*. Wiley, New York.

Nathaniel Beck, Jonathan N. Katz, and Richard Tucker. 1998. Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of PoliticalScience* 42,4: 1260-1288.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag, Heidelberg.

Patrick T. Brandt, John R. Freeman, and Philip A. Schrodt. 2011. Real time, time series forecasting of inter- and intra-state political conflict. *Conflict Management and Peace Science*, 28,1: 41-64.

Bruce Bueno de Mesquita. 2002. *Predicting Politics*. The Ohio State University Press, Columbus, OH.

Nazli Choucri and Thomas W. Robinson, eds. 1979. *Forecasting in International Relations: Theory, Methods, Problems, Prospects.* W.H. Freeman, San Francisco.

Kevin Clarke. 2005. “The Phantom Menace: Omitted Variable Bias in Econometric Research,” *Conflict Management and Peace Science* 22:341-352.

Kevin Clarke. 2012. “More Phantom Than Menace,” *Conflict Management and Peace Science* 29,2:239-241.

John L. Davies and Ted R. Gurr, editors. 1998. *Preventive Measures: Building Risk Assessment and Crisis Early Warning*. Rowman and Littlefield, Lanham, MD.

Robyn M. Dawes. 1979. The robust beauty of improper linear models in decision making. *American Psychologist*, pp. 571-582.

Richard O. Duda, Peter E. Hart, and David G. Stork. 2001. *Pattern Classification*. Wiley, New York, 2nd edition.

The Economist. 2010. Data, data everywhere: A special report on managing information. *The Economist*, 27 February 2010.

Daniel C. Esty, Jack A. Goldstone, Ted R. Gurr, Barbara Harff, Marc Levy, Geoffrey D. Dabelko, Pamela Surko, and Alan N. Unger. 1998. *State Failure Task Force Report: Phase II Findings.* Science Applications International Corporation, McLean, VA.

John Fox. 1991. *Regression Diagnostics: An Introduction.* Beverly Hills: Sage.

Jeff Gill. 1999. The insignificance of null hypothesis significance testing. *Political Research Quarterly*, 52,3: 647-674.

Jack A. Goldstone, Robert Bates, David L. Epstein, Ted Robert Gurr, Michael Lustik, Monty G. Marshall, Jay Ulfelder, and Mark Woodward. 2010. A global model for forecasting political instability. *American Journal of Political Science*, 54,1: 190-208.

Justin Grimmer 2011. An Introduction to Bayesian Inference via Variational Approximations *Political Analysis* 19,1: 32-47.

Jens Hainmueller. 2012. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies *Political Analysis* 20,1 : 25-46.

Carl G. Hempel. 2001. Explanation and prediction by covering laws. In Carl G. Hempel and James

H. Fetzer, eds, *The philosophy of Carl G. Hempel : studies in science, explanation, and rationality*, chapter 5. Oxford University Press, Oxford.

Carl G. Hempel and Paul Oppenheim. 1948. Studies in the logic of explanation. *Philosophy of Science* 15,2: 135-175.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. Bayesian model averaging: a tutorial. *Statistical Science*. 14, 4: 382-417.

Gerald W. Hopple, Stephen J. Andriole, and Amos Freedy, eds. 1984. *National Security Crisis Forecasting and Management.* Westview, Boulder.

Stefano M. Iacus, Gary King and Giuseppe Porro. 2012. Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis* 20,1: 1-24

Daniel Kahneman. 2011. *Thinking Fast and Slow*. Farrar, Straus and Giroux, New York.

Gary King. 1986. How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, 30(3): 666-687.

Gary King and Langche Zeng. 2001. Improving forecasts of state failure. *World Politics* 53,4: 623-658.

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Paul F. Lazarfeld. 1937. Some remarks on typological procedures in social research. *Zietschrift Fuer Sozialforschung* 6:119-139.

Kari Lock and Andrew Gelman. 2010. Bayesian Combination of State Polls and Election Forecasts *Political Analysis* 18,3: 337-348

Barry Markovsky and Evan Fales. 1997. Evaluating heterodox theories. *Social Forces*, 76,2: 511-

525.

John Mearsheimer and Stephen M. Walt. 2013. “Leaving theory behind: What’s wrong with IR scholarship today.” Working paper, Harvard Kennedy School. http://web.hks.harvard.edu/publications/workingpapers/citation.aspx?PubId=8723

Charles A McClelland. 1969. International interaction analysis in the predictive mode. mimeo, University of Southern California, January, 1969.

Sharon Bertsch McGrayne. 2011. *The Theory that Would Not Die: How Bayes’ Rule Cracked the Enigma Code, Hunted Down Russian Submarines and Emerged Triumphant From Two Centuries of Controversy.* New Haven: Yale University Press.

Paul Meehl. 1954. *Clinical and Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press, Minneapolis.

Jacob M. Montgomery and Brendan Nyhan. 2010. Bayesian Model Averaging: Theoretical Developments and Practical Applications. *Political Analysis* 18,2:245-270

Jacob M. Montgomery, Florian M. Hollenbach, and Michael D. Ward. 2012. Improving Predictions Using Ensemble Bayesian Model Averaging. *Political Analysis* 20,3: 271-291.

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

Sean P. O'Brien. 2010. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12,1: 87-104.

John R. Oneal and Bruce Russett. 1999. Assessing the liberal peace with alternative specifications: Trade still reduces conflict. *Journal Of Peace Research*, 36,4: 423-442.

D.W. Orme-Johnson, C.N. Alexander, J.L. Davies, H.M. Chandler, and W.E. Larimore. 1988.

International peace project in the Middle East: The effects of the Maharishi technology of the unified field. *Journal of Conflict Resolution* 32,4: 776-812.

Jon C. Pevehouse and Joshua S. Goldstein. 1999. Serbian compliance or defiance in Kosovo?: statistical analysis and real-time predictions. *Journal of Conflict Resolution* 43,4: 538-546.

Willard Van Orman Quine. 1951. Two dogmas of empiricism. *Philosophical Review* 60: 20-43

Lewis F. Richardson. 1960. *Statistics of Deadly Quarrels*. Quadrangle, Chicago, 1960.

Rummel, R.J. 1979. *Understanding Conflict and War, vol. 4: War, Power and Peace.* Beverly Hills, CA: Sage.

Russett, Bruce M. 1993. *Grasping the Democratic Peace.* Princeton: Princeton University Press.

Russett, Bruce M. and John R. Oneal. 2001. *Triangulating Peace.* New York: Norton.

Gerald Schneider, Nils Petter Gleditsch, and Sabine C. Carey. 2010. Exploring the past, anticipating the future: A symposium. International Studies Review, 12(1).

Philip A. Schrodt. 1990. A Methodological Critique of a Test of the Effects of the Maharishi Technology of the Unified Field. *Journal of Conflict Resolution* 34,4: 745-755.

Philip A. Schrodt. 2006a. Beyond the linear frequentist orthodoxy. *Political Analysis* 14,3: 335-339.

Philip A. Schrodt. 2006b. Forecasting conflict in the Balkans using hidden Markov models. In Robert Trappl, ed, *Programming for Peace: Computer-Aided Methods for International Conflict Resolution and Prevention*, pages 161-184. Kluwer Academic Publishers, Dordrecht, Netherlands.

Philip A. Schrodt. 2009. Reflections on the state of political methodology. *The Political*

Methodologist 17,1: 2-4.

Philip A. Schrodt and Deborah J. Gerner. 2000. Cluster-based early warning indicators for political change in the contemporary Levant. *American Political Science Review* 94,4: 803-817

Science. 2011. “Dealing with Data: Challenges and Opportunities.” Vol. 331 (11 Feb 2011) pp. 692-729

Jasjeet S. Sekhon. 2008. The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods. *Oxford Handbook of Political Methodology*, 271-200.

Philip E. Tetlock. 2005. *Expert Political Judgment*. Princeton: Princeton University Press.

Sergios Theodoridis and Konstantinos Koutroumbas. 2009. *Pattern Recognition*. Springer, Heidelberg, 4th edition

Yaacov Y. I. Vertzberger. 1990. *The World in their Minds: Information Processing,Cognition and Perception in Foreign Policy Decision Making*. Stanford: Stanford University Press.

Jack E. Vincent. 1980. Scientific prediction vs. crystal ball gazing: Can the unknown be known? *International Studies Quarterly* 24:450-454.

Michael D. Ward, Brian D. Greenhill, and Kristin M. Bakke. 2010. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research* 47,5: 363-375.

Nils B. Weidmann and Michael D. Ward. 2010. Predicting Conflict in Space and Time. *Journal of Conflict Resolution* 54,6 : 883-901.

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