Seven Deadly Sins of Contemporary Quantitative Political Analysis*

Philip A. Schrodt
Department of Political Science
Pennsylvania State University
University Park, PA 16802
schrodt@psu.edu

Version 1.0 : August 23, 2010

*Paper 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 and John Freeman, who 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).

Abstract

A combination of technological change, methodological drift and a certain degree of intellectual sloth and sloppiness, particularly with respect to philosophy of science, has allowed contemporary quantitative political methodology to accumulate a series of highly dysfunctional habits that have rendered a great deal of contemporary research more or less scientifically useless. The cure for this is not to reject quantitative methods—and the cure is most certainly not a postmodernist nihilistic rejection of all systematic method—but rather to return to some fundamentals, and take on some hard problems rather than expecting to advance knowledge solely through the ever-increasing application of fast-twitch muscle fibers to computer mice.

In this paper, these "seven deadly sins" are identified as

1. Kitchen sink models that ignore the effects of collinearity;
2. Pre-scientific explanation in the absence of prediction;
3. Reanalyzing the same data sets until they scream;
4. Using complex methods without understanding the underlying assumptions;
5. Interpreting frequentist statistics as if they were Bayesian;
6. Linear statistical monoculture at the expense of alternative structures;
7. Confusing statistical controls and experimental controls.

The answer to these problems is solid, thoughtful, original work driven by an appreciation of both theory and data. Not postmodernism. The paper closes with a review of how we got to this point from the perspective of 17th through 20th century philosophy of science, and provides suggestions for changes in philosophical and pedagogical approaches that might serve to correct 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 still appears on rare occasions, so in the spirit of Samuel Johnson's observation that "Second marriage is the triumph of hope over experience," I'm still reviewing.[1] But the typical paper I receive has some subset—and often as not, the population—of the following irritating characteristics:

- A dozen or so correlated independent variables in a linear model;
- A new whiz-bang and massively complex statistical technique (conveniently available in Stata or R) that is at best completely unnecessary for the problem at hand, since a simple t-test or ANOVA would be quite adequate to extract the few believable results in the data, and not infrequently the technique is completely inappropriate given the characteristics of the data and/or theory;
- Analyses a data set that has been analyzed a thousand or more times before;
- Is 35 ± 5 pages in length,[2] despite producing results that could easily be conveyed in ten or fewer pages (as one finds in the natural sciences);

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.

The issue, when confronting such a paper, is that I do not believe the results. But with the realization 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. Such a paper merely illustrates, alas, what has become "normal science" in quantitative political analysis. "Change the topic, the data, the model, and the interpretation and maybe I'll find this interesting" isn't all that useful. This despite the fact that, fundamentally, there are a lot of things people could be doing very differently[3] and in fact I am thoroughly positive and optimistic about the future prospects for quantitative political analysis. But not the way it is being done now.

The purpose of this paper, which was motivated by an invitation to present on an APSA theme panel titled "A Sea Change in Political Methodology?", is to elaborate on these issues at greater length (and with less individual consequence) than can be done in an article review. As a card-carrying neomedievalist, I will use the trope of "seven deadly sins" to classify these, though I was hard-pressed to limit this to seven: my original list was closer to twenty and a close parsing of this discussion will find those embedded in the larger categories.

[1] Though less so, which will come as a great relief to many reading these remarks.

[2] If 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.

[3] Starting with departments evaluating, rather than counting, articles in tenure files.

This work expands on points I made earlier in Schrodt (2006), and also owes considerable intellectual debt to Achen (2002)[4] and Ward (e.g. Ward, Greenhill and Baake 2010). And in the polemical spirit, perhaps to Huff (1954), which if we can believe Wikipedia, has sold 1.5.-million copies.[5]

To telegraph the conclusion (and to clarify that I am not simply whining about how unfair the universe is to social scientists), I would observe that these problems vary in their solutions but generally fall into the following categories:

- Issues that we know from any introductory methods sequence shouldn't be done but, out of laziness, do anyway: kitchen-sink estimation; analysis employing complex models without assessing whether their assumptions are met;
- Issues where there has been conceptual drift over time: explanation to the exclusion of prediction; Bayesian interpretation of frequentist estimates; interpretation of independent variables as 'controls';
- Issues where technology has made us lazy: endless reanalysis of a small number of data sets using a smaller number of methods;
- Issues that have been debated since the beginning of modern statistics and are probably unresolvable: frequentism more generally; the interpretation of significance tests from population data;
- Issues that are simply hard but where some additional work—intellectual as well as technical—might move the field forward: a better philosophy of statistical inference based on current Bayesian practice and incorporating this, rather than mid-20th century frequentism and logical positivism, into our methodological pedagogy.

Finally I would note that there is a certain bias in my discussion towards the fields in which I am most likely to review, the quantitative analysis of political conflict in both international and comparative forms. The issues I am raising may be less relevant to, say, studies of United States voting behavior. Though I doubt it. Readers with sensitive constitutions may also find the paper at least mildly polemical. I don't doubt that. "Being on an APSA roundtable means never having to say you're sorry."[6]

1 Kitchen-sink models and the problem of collinearity

Kitchen-sink models—Achen prefers the term "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." This point is listed as the first of the deadly sins not because I plan to expound on

[4]Who makes about two-thirds of the points I want to make here, says them better than I have, and said them ten years ago. And echoes a number of the same points King (1986) made 15 years before that. And a lot of good all of this seems to have done us... so... I'll say'em again...

[5]http://en.wikipedia.org/wiki/How_to_Lie_with_Statistics. Accessed 21 August 2010.

[6]http://www.imdb.com/title/tt0066011/

it in detail, nor because it hasn't been said before—Achen (2002) has reassuredly been *cited* 166 times according to Google Scholar[7]—but rather because it is the single most frequent offender and the source of maybe 80% of my distrust of contemporary quantitative research.

Achen's succinct "Rule of Three" asserts (backed up with an assortment of methodological and technical justifications)

> 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)

My own spin on this (Schrodt 2006) focuses on the most common context where the issue is encountered, the serious pathologies that quickly emerge when correlated variables are used in regression or logistic models.[8] 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 orthogonally, 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.

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 omitted variable bias problem—and its explanatory power is reflected in whatever related variables happen to be in the equation. Try to include any but the simplest categorical variables and the analysis is nibbled to death by dummy variables (whose coefficients, in practice, are only rarely interpreted correctly, at least in articles I review).

In the absence of a strong linear effect in the main population, regression *amplifies* rather than isolates the influence of anomalous subpopulations. How many published statistical results are in fact 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 thoroughly check for this possibility.

In short, for many if not most 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—take the interminable democratic peace debate in international relations (please...)—but not so great at ever coming to a conclusion. The orthodox response to this: "Well, you have to resolve these inconsistencies on the basis of theory." But usually the

[7]Now, folks out there in journal-land, will you please follow the article's admonitions, not just cite it?

[8]The remainder of this section is heavily auto-plagiarized from Schrodt 2006.

whole point of doing the test in the first place was to differentiate empirically among 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.

2 Pre-scientific explanation in the absence of prediction

I was recently sent the draft of the introduction to an edited journal issue in which I had an article dealing with prediction. The draft—I believe it was eventually changed—began with an all-too-familiar sentiment in my part of the discipline: "Many international relations scholars still see prediction as an inferior task in comparison to explanation."

The "'Many international relations scholars still see" segment is incontestable: this is a deeply-held article of faith among a sizeable segment of that community. It's the "prediction as an inferior task in comparison to explanation" part I have problems with.

My subtle and finely nuanced assessment of that sentiment: This is utterly, totally and completely self-serving bullshit, devoid of even the slightest philosophical justification, tremendously damaging to our collective intellectual enterprise, and best dispatched to the trash compactor of history to be revived only as an example of what *not* to do.

Have I made myself clear?

So, what's the big deal here?... like, just kick back, dude. Okay, okay, bit of a sore spot since I continually get hit by this and yet not once, *not once*, has anyone been able to provide even a single citation in the philosophy of science literature to justify it. Instead, it is simply folk wisdom derived, I suspect, from the following syllogism

> The models we are working with are nearly worthless for prediction
>
> We are valiant scientists of unparalleled intellect and vision
>
> *therefore...*
>
> Scientific models do not need to predict

A perfectly understandable human impulse, of course—if you can't make the goal, 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 (Goldstone et al 2010) and the Integrated Conflict Early Warning System (O'Brien 2010) being conspicuous recent examples, even if most models don't do this (Ward, Greenhill and Baake 2010). Furthermore, this is certainly not where the quantitative international relations community started—the early proponents were decidedly interested in developing models that were predictively accurate as they were motivated in part by the hopes of applying such knowledge to reduce the probability of getting their butts fried in a US-Soviet thermonuclear conflagration. But things have drifted since then.

While the contemporary sentiment is understandable, it is philosophically vacuous. Again, the most exasperating aspect of this assertion is that its proponents—at least in my experience—simply repeat it, as though repetition will somehow make it true. So in the absence of a

specific argument to refute, one can only provide evidence to indicate the contrary. And even this is somewhat difficult because of the complete disconnect between this sentiment and 20th century philosophy of science: That was, for the most part, coming out of the deterministic predictive traditions of Newton, Laplace, Maxwell and pretty much the whole of the experience of the natural sciences since the mid-17th century, so the centrality of prediction—not just explanation—in scientific practice is simply taken for granted in most discussions.

Nonetheless, with a bit of digging one can find some succinct arguments. For example, the fact that Carl Hempel's the classic covering law essay is titled "Explanation and Prediction by Covering Laws" (Hempel 2001) would suggest that for Hempel the two go together in a properly scientific theory. The text of that essay as well as the title supports this contention: throughout the essay Hempel is unambiguously treating explanation and prediction as equivalent. The logical positivists, being rather rigorous logicians (sometimes maddeningly so[9]), would of course be completely appalled at the notion that two things could simultaneously be equivalent and one of them weaker than the other: accept that premise and one can derive no end of absurdities.

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

> Let us note here that the same formal analysis, including the four necessary conditions, applies to scientific prediction as well as to explanation. The difference between the two is of a pragmatic character. If E is given, i.e. if we know that the phenomenon described by E has occurred, and a suitable set of statements $C_1, C_2..., C_k, L_1, L_2, ..., L_i$, is provided afterwards, we speak of an explanation of the phenomenon in question. If the latter statements are given and E is derived prior to the occurrence of the phenomenon it describes, we speak of a prediction. It may be said, therefore that an explanation of a particular event is not fully adequate unless it's 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 while the car was traveling at high speed. 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, nor does it state adequately the antecedent conditions which would be needed for the prediction." (Hempel and Oppenheim (1948: 138-139)

This is the critical insight from Hempel (and the logical positivists more generally): explanation in the absence of prediction is not somehow scientifically superior to predictive

[9]Consider, for example, the effort expended by the logical positivists in trying—unsuccessfully—to resolve the conundrum that, logically, a yellow pencil is evidence for the statement "All crows are black" by virtue of the fact that "All crows are black" and "All non-black things are non-crows" are logically indistinguishable.

analysis, it isn't scientific at all! It is, instead, "pre-scientific." Hempel's contemporary Quine, quoted in the final section, will make much the same point.

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 Northern Europeans—dead white ax-wielding European males—was that lightning bolts were hurled by the Norse god Thor. For the 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. And a decided improvement on reading the mind of Thor. Or sacrificing goats to Thor in order to protect buildings. Sven sacrifices goat to Thor; Sven's barn burns down. Helga installs new fangled lightning rod; Helga's barn survives. Electricity theory good; Thor theory not so good.

Forward to the 20th century, and theories combined with experiments on the artificial generation of lightning by Tesla, van de Graff and others, and increased empirical understanding—through a combination of basic theory progressively refined through observation, [still hazardous] experimentation and simulation—of the conditions giving rise to lightning, and one sees, for example, mathematical models being used to predict where lightning is most likely to ignite forest fires, and firefighting resources allocated accordingly. Sometimes the predictions are incorrect—weather systems are complex, and the effects of an individual lightning strike even more so—but still, this a great improvement over trying to read the mind of Thor.

This analogy is not a mere cheap shot: I actually found the quote from Quine (1951) that I use in the final section—Quine prefers Greek mythology—well after I had written this example, and whether this is a case of convergent reasoning or my associative memory dragging forth some long-buried lore from John Gillespie's research and methods course I took in graduate school, the point is that distinguishing scientific explanation from mythical (or other non-scientific, such as Freudian) explanation is one of the central themes for the logical positivists. In the absence of prediction, it cannot be done.

There is, of course, a place for pre-scientific reasoning. Astrology[10] provided part of the

[10]Again, this comparison is anything but a cheap shot: astrology has virtually all of the components of

empirical foundation for astronomy, alchemy the same for chemistry, and no less a scientific mind than Newton devoted a great deal of attention to alchemy.[11] Furthermore to my pluralist mindset, the scientific mode of generating knowledge is not the only valid way that we can learn about politics, and the 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, and merely calling them scientific does not make them so. Only prediction will do this.

3 "Insanity is doing the same thing over and over again but expecting different results."[12]

When things become very easy, they either revolutionize our lives (transportation 1820-1960, antibiotics, the internet and cell phones), or allow us do things that are very stupid (World War I/II, television, Twitter). More to the point, science will advance when relatively routinized procedures that can be implemented systematically by a large number of people lead to incremental advances in knowledge, but stalls when the increments to knowledge that can be obtained routinely have been exhausted.

We are presently in such a stall: many of the easy things have been done, and routinized procedures often now only further contribute to confusion, because any finding can be undone by a slightly different analysis of the same data, and even an expert (to say nothing of the general public) will have a difficult time telling these apart. Consequently, I believe very little of what I'm reading in the journals, which 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, and Oneal and Russett's (1999) original analysis of the Oneal-Russett data set on the democratic peace was interesting. Quite possibly, Oneal and Russett missed something really important and a few other articles using their data set would be worthwhile. But 113 articles?[13] Particularly when most of those articles are just minor specification, operationalization or methodological variations on the original, collinearity-fraught data set? What, other than essentially random fluctuations in the coefficient values and standard errors, are we going to get out of this?

Not all of those citations, of course, involve a reanalysis of the data. Just too many. Let's assume, conservatively, that only 50% involve re-analysis. Let's also assume—this may or may not be accurate—a 1-in-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

a legitimate scientific enterprise *except* for predictive validity, and the challenge of differentiating astrology from orthodox science has been an issue in philosophy of science since the time of Bacon.

[11] As a result, if we are to believe Wikipedia, "John Maynard Keynes, who acquired many of Newton's writings on alchemy, stated that 'Newton was not the first of the age of reason: He was the last of the magicians.' " http://en.wikipedia.org/wiki/Isaac_Newton, accessed 17 Aug 2010.

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

[13] The current ISI count of the cites to the article.

to frequentist practice—that the data have been re-analyzed about 3,000 times.

And apparently, compared to fields studying the American National Election Survey, we in IR should consider ourselves fortunate.

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, though formalizing a measure of this is surprisingly difficult, a finite amount of information in any data set (as well as modifications of the data with covarying indicators). When competently done—as is usually the case when someone has invested the effort in collecting, rather than merely downloading, the data—the first five or ten articles will figure out those effects. But I defy the reader to provide me 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 difference. Methods robust to the existence of collinearity such as cluster analysis and principal components will just ignore this—they've already detected the relevant latent dimensions in the existing indicators. Brittle methods—regression and logistic—will merely go berserk and rearrange the coefficients based on subtle interactions (sometimes, in fact, 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 is the number of data sets that are insufficiently analyzed as a consequence. Systematic data collection is something we *really* have down now: The number of well-documented and reasonably thoroughly collected (nothing is perfect, and certainly not the data sets I've generated...) data sets now available is astonishing, and nothing remotely resembling this situation existed even 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. We're doing the science right on this aspect of methodology.

Furthermore, data collection is *not* a monoculture: we should be in a situation where we can get similar results from multiple convergent indicators. But, for the most part, this is not happening due to a focus on 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*), and the fact that convergent indicators are toxic in a regression framework. On this issue I believe the answers are hiding in plain sight—we have robust methods, and certainly no one is forcing people to reanalyze the same data sets over and over and over again—and many of the reasons those alternatives are not pursued are due to historical idiosyncracies (e.g. the earlier mis-use of data reduction methods such as factor analysis and LISREL) that could easily be corrected.

4 Using complex methods without understanding the underlying assumptions

About eight months ago I was sent an article to review using a methodology I had been previously unfamiliar with, competing risk models. The article confidently cited the original work which derived the test, so I went back and looked at it, and I also did the requisite Web searches, and quickly established that a core assumption for the estimation was that "Each failure mechanism leading to a particular type of failure proceeds *independently* of every other one, at least until a failure occurs." (http://itl.nist.gov/div898/handbook/apr/section1/apr181.htm; emphasis added). In the paper, the failure modes studied were not even remotely independent—as I recall they were pretty much nested—so I wrote a relatively brief review noting this, indicating that as a consequence the results were probably meaningless, and figured that would be the last I would hear of competing risk models.

Alas, no. About three months later a practice job talk presentation by one of our very best graduate students featured a competing risk model. With risk factors that were correlated both theoretically and, I presume, empirically. I raised the issue of the violation of the assumptions of the model and received the somewhat embarrassed response "Well, yes, I know, but everyone else does this as well."[14]

This is mouse-clicking, not statistical research. Competing risk models may in fact be robust against violations of the assumption of independence. However, this is certainly not in the original article—where the results depend heavily on the assumption of independence, and the derivations would be considerably more complex, if not completely intractable, without that assumption—and if such a result exists elsewhere in the literature (which it might) it was not cited in either the paper or presentation. But it is quite likely that the statistical properties which recommend competing risk models in the first place are weakened by violations of the assumption of independence. Not infrequently, complex methods whose assumptions are violated will perform worse than simpler methods with more robust assumptions.

There is nothing unique about the competing risk model in this regard: over the years, I have seen countless examples where a paper uses a complex method—as with competing risk models, 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.[15] Often as not, "everyone is doing it" carries the day with the journal editors—methodologists are such lame and boring nags after all, worrying about fundamentals and all that other useless picky stuff. This will set off a cascade of equally bad papers—after all, doing things badly is usually easier than doing them correctly—until someone notices that most of the resulting estimates within that literature

[14] He got the job anyway.

[15] To cite another personally irritating case, in the 1990s there was an obsession, lasting about a decade, in using the Augmented Dickey-Fuller (ADF) test to assess the possibility of co-integration in time series models using event data. This was problematic in two regards. First, the ADF has notoriously low statistical power—approaching zero—in data that are highly autocorrelated but still stationary ($\rho <\approx 1$ in $y_{t+1} = \rho y_t$), which describes a lot of event data. But more importantly, the data generation process for event data *guarantees* it is bounded (and hence, in a long time series, stationary in the sense of the co-integrative alternatives), so there was no point in doing the—in all likelihood misleading—ADF test in the first place.

are incoherent and might as well have been produced by a random number generator, and we sidle off to the next set of mistakes.

Once again, Achen (2002) has covered this ground rather thoroughly. And fruitlessly, given my experience. 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, Katz and Tucker (1998). Quantitative IR was analyzing a lot of binary time-series cross-sectional 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 the 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, where are numerous cases where one sees complexity for the sake of complexity, in the hopes (often, alas, realized) that using the latest whiz-bang 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 the future there will be a big party when you get tenure. But, in fact, the whiz-bang technique probably makes at best marginal changes to your coefficient estimates and standard errors because it is only effective if you know things you don't know—such as the variance-covariance matrix of the errors in your equations, or the true propensity function in a matching problem. Or—seen with increasing frequency— the whiz-bang method as you have actually applied it reduces to something much simpler, as in "How many ways can you spell F-L-A-T- -P-R-I-O-R-S?"

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'll be decidedly more willing to believe your results because the t-test is robust and requires few ancillary assumptions (and the key one is usually provided, for free and indisputably, 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 yoga master 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 at least you will be able to interpret the OLS coefficients).

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

5 If the data are talking to you, you are a Bayesian

At the pedagogical and mainstream journal level in political science—though no longer at the elite level—we have legitimated a set of rather idiosyncratic and at times downright counter-intuitive 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), 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 it is that in fact it doesn'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). To call attention to only the most frequent [sic] of these problems as they are encountered in political science:

- 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. Which is not what you want to know—you usually want to know the magnitude of the effect of an independent variable, given the data. That's a Bayesian question, not a frequentist question. Instead we see—constantly—the p-value interpreted as if it gave the strength of association: this is the ubiquitous Mystical Cult of the Stars and P-Values which permeates our journals.[16] This is not what the p-value says, nor will it ever.

  In my experience, this mistake is almost impossible to avoid: even very careful analysts who are fully aware of the problem will often switch modes when verbally discussing their results, even if they've avoided the problem in a written exposition. And let's not even speculate on the thousands of hours and gallons of ink we've expended correcting this in graduate papers.

- The frequentist paradigm—leave aside the internal contradictions with which we have somehow coped for close to a century—does apply fairly well in the two circumstances for which it was originally developed: random samples and true experiments. These situations apply 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, and never will: pretty much the whole of IR, comparative research except that involving surveys, and most of national political studies except for those involving public opinion (e.g. almost all studies of executive, legislative and judicial behavior).

  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

[16] Which supplanted the earlier—and equally pervasive—Cult of the Highest R^2, demolished, like Thor's oak at Fritzlar, by King (1986).

methods on populations, they are fundamentally different. A debate that has a very long history: see Morrison and Henkel 1970.

- The ease of exploratory statistical computation has rendered the traditional frequentist significance test all but meaningless.[17] 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. In fact the sheer information management requirements of adjusting for the 3,000+ 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, or...) render such an adjustment impossible.[18]

- Finally—well, 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, and so rejection of the null hypothesis merely confirms this.

  Granted, if one is operating in a strict falsification framework—and somehow can get around measurement and collinearity problems, the inaccuracy of the significance test in the estimation of multiple closely related specification and so forth, and actually believe the results rather than trusting your instincts (that's Bayesian again!) and estimating yet another alternative formulation of the model—acceptance of the null hypothesis is useful. And if in numerous alternative formulations the variable still isn't significant, that is probably fairly good evidence to reject its relevance, and that is progress.

  But as a long literature[19] has established—this was one of the jumping-off points for Kuhn—scientific inquiry, while accepting the *principle* of falsification, only rarely proceeds using strict falsification norms. Instead, the general tendency is to do extensive exploratory work and substitute paradigms only when a superior alternative is available (Lakatos; this is also very closely related to the problem of ancillary assumptions addressed by Quine and Duhem). To the extent that we are not working in a strict falsification framework—and in the stochastic realm of social behavior, this is pretty darn

[17] A situation dramatically different from that as recently as fifty years ago. My father obtained his doctorate, in education, fairly late in life—albeit writing a dissertation on what I still find to be one of the most clever dissertation topics ever, "Why people don't finish their dissertations"—and I recall as a teenager listening in the mid-1960s as he and his fellow students in the required statistics course complained about a particular homework assignment that was notorious for taking days of tedious calculations (in our household, actually done by my mother). This was, I am almost certain, a multiple regression, probably with a sample size well under 100 and only three or four independent variables.

[18] This paragraph largely auto-plagiarized from Schrodt 2006.

[19] Which I can provide at a later date but this paper is already *way* too long.

close to "always"—the failure to reject a null hypothesis in a single instance (which nominally is how the frequentist approach is supposed to work) is telling us almost nothing.

The alternative to these problems is, of course, Bayesian approaches. At the elite level these are already widely accepted: at least fifteen years ago some august member of the Society of Political Methodology[20] intoned "We are all Bayesians now." But, as I will discuss in more detail later, this has not filtered down to either our instruction nor, generally, the practice in the discipline as a whole. For example Bayesian approaches are common in the consistently top-ranked (by ISI citation counts) *Political Analysis* but not in more mass-market venues such as the *APSR* and *AJPS*, which are overwhelmingly frequentist.

Yet behind every poor graduate student or assistant professor who is struggling to figure out why it makes any sense to do a null hypothesis test on a population (hint: it doesn't...), there is a Bayesian struggling to get free. Free our inner Bayesians, and we also solve a series of additional problems regarding evidence, inference and applications. These issues will be pursued in more detail below.

6 Enough already with the linear models!

Even the most cursory glance at quantitative studies in the mainstream journals over the past twenty years[21] will show that—the recent applications of Bayesian methods aside—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 data in real-world applications that, in fact, plots out nicely as a line, perhaps with a few interesting and plausible outliers. Elite media—a.k.a. *The Economist*—have even been known to include correlation coefficients in the occasional graph.

But, as any good'ol'boy teaching at a farm school[22] can tell you, monocultures always have the same unhappy ending: parasitism and disease,[23] followed by collapse. Which I suppose is another way of characterizing the entire point of this essay.

[20] probably Larry Bartels, who is particularly adept at intoning memorable phrases.

[21] Only: prior to that one also encountered quite a few nonparametric statistics applied to contingency tables, and even the occasional ANOVA and variants on factor analysis and other latent variable models, some actually done correctly. Though ANOVA, factor analysis and LISREL are linear.

[22] Hint: they usually end in "State"

[23] Parasitism in this context is the individual who, year after year, grinds out articles by downloading a data set, knocks out a paper over the weekend by running a variety of specifications until—as will invaribly 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, will eventually publish this. And who believes this to be legitimate "normal science" and makes well known their sense that anyone who operates otherwise is a naïve chump. Dear reader, does this characterization not sound familiar? Do we not all know at least one person fitting this description? Yet another of the dirty little secrets of frequentism.

The problems with this particular monoculture have been detailed elsewhere in this essay; the point I want to make in this section is that there are alternatives. Consistent with my monoculture metaphor—consult your local "heirloom gardening" fanatic—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 *The Economist* (2010) and Schrodt (2009). 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 inference.

The number of methods we are *not* using is stunning. In 2000, a number of U.S. political methodologists attended the International Sociological Association's methodology section meeting in Cologne, Germany. For starters, we were surprised at how huge it was—sociology has developed a bit differently in the lands of Marx, Durkheim, and Weber than it has on this side of the pond—but also in the large number of highly sophisticated studies using correspondence analysis (CA), a method almost unseen on this side of the Atlantic. Yet CA 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 (there are currently four different implementations in *R*, as well as open-source code in almost all current programming languages), and yet generally absent in political analysis except in applications to natural language processing.

Finally, in the purely qualitative realm—generally *terra incognito* to regression-oriented researchers—the machine learning community has developed a wide variety of classification tree algorithms for categorical data, ranging from the early ID3 and C4.5 methods to contemporary variants such as CART and CHAID. Again, robust, well-understood, readily-available in open-source code, and essentially invisible in the political science community.

This is the tip of the iceberg. Just sampling from three current texts on computational pattern recognition (Duda, Hart and Stork 2001, Bishop 2006, Theodoridis and Koutroumbas 2009), we find in addition to the methods discussed above

- multiple variations on neural networks
- multiple variations on Fourier analysis
- multiple variations on principal components
- hidden Markov models
- sequential, functional, topological and hierarchical clustering algorithms
- multiple variations on latent variable models

- genetic algorithms and simulated annealing methods

The issue with these alternative methods is not just novelty-for-the-sake-of-novelty—this would be as dysfunctional as the complexity-for-the-sake-of-complexity I criticized above. It is rather that at a minimum these techniques employ alternative structures for determining regularities in data—just because lots of things are linear doesn't mean that *everything* is linear—and in many cases, they deal with issues that are commonly found in political science data. SVM and decision-tree methods, for example, 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 those 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, this is almost all situations).

On this matter (for the first time in this essay), I am actually somewhat optimistic, and much of that optimism can be summed up in a single letter: *R*. The *R* statistical package has become the *lingua franca* for systematic data analysis in all fields—incongruously, for example, there are complete text analysis packages in *R*—which has broken down past disciplinary barriers that resulted from the understandable tendency of commercial statistical packages to specialize by applications. The open-source *R* has promoted the rapid diffusion of new methods: We are rapidly reaching a point where any new published statistical method will be accompanied by an implementation in *R*, at least available from the author, and if the method attains any sort of acceptance, from CRAN. And finally, the current version of *R* is just a few mouse-clicks away, and one needn't worry about the departmental computer lab having only Version 8 when the routine you want to try is in Version 11.[24]

The availability of a method, of course, is no guarantee that it will be properly used, and I think some changes will be needed in our pedagogy. At the moment we tend to educate methodologists to nearly the limits of the linear model, usually through advanced econometrics texts such as Gajarati or Maddala. Yet such a student would find even the most basic application of CA, SVM or CART completely unfamiliar.[25] My guess—though this is going to take a lot of experimentation—is that we will eventually need to move to a pedagogical approach that emphasizes more basics, and robust methods from a variety of fields (yes, even ANOVA), at the expense of the intensive drill-down into a single methodological approach that we currently use.

[24]The downside of R is its complexity and, well, weirdness: I have heard on multiple occasions, at multiple institutions, usually from individuals who use, but do not develop, statistical methods, "Our graduate students are no longer learning statistics, they are spending all of their time learning *R* programming."

I can sympathize with this: as a *programmer* I find *R* to be a major advance over prior alternatives, notably `SAS` and `Stata`. But *R* requires thinking in a rather peculiar manner, and a manner that is, in fact, closer to programming than it is to statistics. `SAS`, `Stata` or, in an earlier era, `SPSS`, were much closer to the technique and for a quick analysis, I—like many users of *R*—will still use `Stata`.

In the long term I think *R* will still win out due to the fact it is open-source, and in the long run we will probably get a user-friendly interface on top of *R*, much as the Macintosh OS-X and Ubuntu Linux put a user-friendly interface on top of Unix. The self-identified "power user" will use the command line; the casual user need not even know that the command line exists. But despite some efforts, we're not there yet.

[25]Having mastered advanced econometrics, a student should be able to figure out those methods, but the student would not have encountered them in the classroom nor experimented with them.

7 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).[26] 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 a situation where, again, I think 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—and having clear historical antecedents—much of the original frequentist work was in the context of controlled experiments—is, well, simply dead wrong in the contexts in which we apply it today, estimation of linear coefficients from sets of correlated independent variables measured across inhomogeneous populations. We know this is wrong but we forget it is wrong, in an manner similar to interpreting a p-value as the measurement of the magnitude of the effect of an independent variable.

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 class a cross-national data set and have them derive the most ludicrous model they could find in terms of obtaining significant coefficients on nonsensical independent variables as a result of spurious correlation or, more commonly, collinearity effects. None had the slightest problem doing this. Fortunately 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 other side of this coin—though this could as well have gone under the category of pathologies of frequentism[27]—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

[26]Since I will later invoke the Church of Scientology, a new religious movement notoriously adverse to criticism, I would like to note that in my twenty years of subsequent interactions with the TM community, they have, without exception, been unfailingly cordial, courteous and professional.

[27]As I said, there are actually about twenty criticisms in this paper, not just seven.

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 from messy data from a population.

This has serious implications. Much of the early work of the Political Instability Task Force, for example, 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, Greenhill and Bakke 2010). Only when PITF methodology shifted to modes of assessment that specifically considered 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.

And prediction, not explanation, is what establishes a study as scientific.

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.[28] Given the number of well-studied pathologies in human intuitive reasoning (Vertzberger 1990, Tetlock 2005), even among experts, we need all the help we can get to figure out political behavior. Instead, I suggest that we take seriously these criticisms as the initial 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.

As I noted in the first section, a number of my criticisms involve nothing more than getting out of some bad habits that we've always known we shouldn't be doing: Methodology 101 material. Kitchen-sink models are meaningless; significance tests tell you nothing more than something you probably didn't need to be told in the first place; 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 the same data—there's probably a serious need to go back and clean

[28]To elaborate: My concern is that this paper will be—to use the current socially-constructed meme (life span $\approx$ that of tofu)—"Shirley Sherroded" (Google it...): quoted out of context to support precisely the opposite point I am making. So, Schrodt, Schrodt!, The Former President of the Society for Political Methodology Schrodt, Echoing Former President of the Society for Political Methodology Achen, Says Most Quantitative Research is Crap!!! Dust off those copies of Foucault, Derrida and the *Oxford Encyclopedia of Snarky Phrases*, throw the computers out the window and score some really good drugs as your primary research tool, 'cuz postmodernism is a-coming back into vogue, whoopee! No, that's not really my point here.

More seriously, on a number of occasions I have been scolded and admonished that I shouldn't say "this stuff" because "they" will misuse it. But "they" are going to do whatever "they" want irrespective of what we do or do not say, nor will "they" even read what we say except possibly if it has a clever title like *Seven Deadly Sins* and is on an APSA server, and in the meantime letting the folks who could be doing things better run into various swamps for fear that our enterprise is so fragile that attacking *anything* is attacking *everything* (which it isn't) is not a particularly constructive strategy. IMHO.

up the mess, and some of that mess is in a larger professional context (e.g. lazy tenure committees[29]). Finally, while technology has been a curse in some regards, it clearly can assist in the transition from frequentist to Bayesian estimation, as well as opening up a variety of new techniques, many of which are already in use in applied settings.

Beyond this list, however, there are some things that need to be done where we still don't have clear solutions, where the intellectual groundwork still has yet to be completed. And furthermore, some of this work is likely to be hard, and—gasp—the solution does not lie in the further application of fast-twitch muscle fibers.[30] An outline of part of that agenda circuitously follows.

I will start by stepping back and taking a [decidedly] bird's eye (Thor's eye?) view of where we are in terms of the philosophy of science that lies beneath the quantitative analysis agenda, in the hope that knowing how we got here will help to point the way forward. In a nutshell, I think we are currently stuck with an incomplete philosophical framework inherited (along with a lot of useful ideas) from the logical positivists, combined with a philosophically incoherent approach adopted from frequentism. The way out is a combination of renewing interest in the logical positivist agenda, with suitable updating for 21st century understandings of stochastic approaches, and with a focus on the social sciences more generally. Much of this work has been done last decade or so in the qualitative and multi-methods community but not, curiously, in the quantitative community. The quantitative community does, however, provide quite unambiguously the Bayesian alternative to frequentism, which in turn solves most of the current contradictions in frequentism which we somehow—believing six impossible things before breakfast—persuade our students are not contradictions. But we need to systematically incorporate the Bayesian approach into our pedagogy. In short, we may be in a swamp at the moment, but the way out is relatively clear.

How we got here

How *did* we get here??... again, the Thor's-eye view. Coming out of a rather abysmal millenium or so of hyper-deductive intellectual paralysis that culminated in Scholasticism,[31] something reasonably close to the modern scientific method is developed in the late Renaissance in a fairly coherent set of developments initiated by Bacon around 1600 and pretty much wrapped up by Descartes by around 1640. This was sufficient to lay the groundwork for the 17th century's most vaunted scientific accomplishment—Newton's mechanics—as well as the Englightenment. We move through the 18th century with few challenges to this approach that are still relevant today—I'm telling the story from the perspective of the winners—except possibly the radical skepticism of Hume. By the 19th century we are encountering recognizably modern efforts to apply these methods to the social sciences—Mill,

[29]There is a serious move afoot, long enforced in NSF proposals, that tenure should be based on your N best articles—where N is generally in the range of 5 to 10—which would certainly go a long way to removing the incentives for the proliferation of analyses which everyone, including the author, is fully aware are redundant and at best produce a marginal contribution.

[30]except possibly to http://plato.stanford.edu/

[31]which, it is frightening to note under the circumstances, survived in the academy a good century or two after it was delegitimated in more pragmatic sectors of society. When Charles I was confronted by a rather well-armed segment of the new English middle class who were unimpressed by the doctrine of the divine right of kings, he fled to the more congenial, if not secure, precincts of Oxford.

Bentham, Marx, Pareto—though with much greater success in applying rigorous philosophical criteria (e.g. James, Mach) and the experimental method (Wundt, Pierce, Dewey) in the behavioral sciences (modern psychology).

Into this situation came the Vienna Circle of logical positivists in the early 20th century. At this point, it is useful to review what the logical positivists did and did not have available to them, as well as where they were situated scientifically and philosophically.

First, the overall objective of the logical positivists was a systematization of the foundations of scientific inquiry that would parallel that achieved in mathematics (notably Cantor, Hilbert and culminating in the work of Russell and Whitehead), and also in several breakthroughs in unifying physical laws, notably Maxwell's equations. In addition, responding to the glorification of "science" since the Enlightenment, and certainly accelerating with the Industrial Revolution, they were consciously seeking to differentiate orthodox science from various challenges, most notably Freudian psychology and political Marxism.[32]

A great deal of what the logical positivists systematized (some from much earlier sources such as Bacon and Descartes, other more recent) is material that, I believe, we are definitely going to keep—behavioralist concepts of measurement, the concept of experimentation, falsification, the deductive/inductive distinction and so forth: in general, the first three chapters of King, Keohane and Verba (1994; KKV).[33] However, it is just as important to note two critical things that, from our 21st century perspective, the logical positivists did *not* have.

First, and probably most critically, they did not have a good understanding of probabilistic mechanisms and reasoning. Statistics was still relatively new (and often as not, marginalized as an applied, rather than theoretical, field, a situation that would not change in many universities until late in the 20th century, if then). They were instead working primarily in the context of deterministic systems such as Newton's and Maxwell's, and—as the well-studied early 20th century discomfort with quantum mechanics shows—generally found non-deterministic systems rather weird. This focus on determinism also allowed them to deploy the full strength of logic to the problem (again, paralleling Russell and Whitehead in mathematics).

Second, the *original* objective reached a dead-end—or perhaps more accurately, its logical limits—in the 1950s with Carnap, Quine, Duhem and others who established that the synthetical/analytic distinction was not going to be established on a purely logical grounds, due largely to the problem of the infinite regress of ancillary assumptions in a theory.[34] Which, indirectly, led us to the current prediction/explanation mess: it is interesting to note

[32]These were the two challenges which gained the most attention of the Vienna Circle, but hardly the only to claim to share the mantle of "science." To take one of the most successful, Mary Baker Eddy choose to title her highly successful late 19th century faith-healing movement "Christian Science." Need I also remind us that the American Political *Science* Association was founded in this same era? Said moniker having preceded by at least two decades the existence of any genuinely scientific work in the field, that of Charles Merriam and his collaborators at the University of Chicago in the 1920s.

The trend continues: Mary Baker Eddy's successful marketing gambit was emulated in the 1950s by the aspiring science fiction writer L. Ron Hubbard who would create a space-age melange of 2000-year-old Gnostic theology and Freudian psychoanalysis and title it... the envelope, please... "Scientology."

[33]Logical positivist critiques of assorted German philosophers, from Hegel to Heidegger, might also still have some utility. Particularly Heidegger.

[34]Distinct from, but similar to, Hume's earlier issues with infinite regress in establishing the empirical validity of induction: "turtles all the way down" as the current meme goes.

that Quine phrases this as

> As an empiricist I continue to think of the conceptual scheme of science as a tool, ultimately, for predicting future experience in the light of past experience. Physical objects are conceptually imported into the situation as convenient intermediaries not by definition in terms of experience, but simply as irreducible posits comparable, epistemologically, to the gods of Homer . . . For my part I do, qua lay physicist, believe in physical objects and not in Homer's gods; and I consider it a scientific error to believe otherwise. But in point of epistemological footing, the physical objects and the gods differ only in degree and not in kind. Both sorts of entities enter our conceptions only as cultural posits. (Quine 1951)

Note—please, please—that Quine is not saying that it is "scientific" to be reading the minds of the gods—Quine as a *scientist* believes in predictive models. But due to the problem of ancillary assumptions, Quine as a *logician* cannot establish this. A similar problem, albeit from completely different sources—a rather quirky little result in number theory, Gödel's incompleteness theorem—had arisen about twenty years earlier to limit work on the foundational approaches to mathematics, and by the 1950s quantum mechanics had become well-established as an alternative to Newtonian/Leplacian determinism. Thus, arguably, the philosophy of science community was ready for the original agenda to end [somewhat] unhappily in this manner.

So just at the time when, more or less coincidentally, modern political science behavioralism was beginning to take off—largely as a consequence of technological developments in computing and the influx of a whole lot of money during the Cold War expansion of higher education in the U.S.—the logical positivists are leaving the field. Subsequent "philosophy of science" in the remainder of the 20th century takes a sociological-historical approach, beginning with Kuhn and Lakatos, then on through the postmodernist scientific culture wars of the 1990s (e.g. Sokal 1998) and into the relative vacuum—at least as it appears in our texts—of today. As of the early 21st century, we're in a situation comparable to the end of the climactic car chase scene in a movie: the postmodernist car has gone off the cliff, and is now sitting at the bottom in a burning heap, and the bad guys are [philosophically] dead,[35] and now what we need to do is go back and wrap up all of the loose ends in the plot that were left unresolved while we dealt with the bad guys.

Frequentism can be disposed of much readily, since frequentism—whatever its practical utility—never made any logical sense in the first place, which probably goes a long way to explaining both why the logically rigorous Vienna Circle paid little attention to the nascent philosophy of statistical inquiry, and why we batter our heads against proverbial walls every fall semester trying to persuade graduate students that frequentism makes sense when it doesn't.

Frequentism as applied to the social sciences suffers from three flaws, all potentially fatal. First, aforementioned inconsistencies of the Fisher-Neyman-Pearson ABBA compromise. Parts of the practice—notably confidence intervals applied to random samples—still have a great deal of utility (and one need not apologize to graduate students about teaching confidence intervals, or distributions, or the Central Limit Theorem, or probability theory, or

[35] If they were ever [philosophically] alive: insert obligatory zombie and vampire references here...

even the fact that linear combinations of normally distributed variables are linear... again, don't throw out the proverbial baby with the proverbial bathwater. Or dishwater.) But the frequentist *package as a whole* does not make logical sense, and while we've managed to work around this for the better part of a century, it does not provide a foundation for going any further, and we may have hit just those limits.

Second—and perhaps more important to us as political scientists—one simply cannot juxtapose 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 pretty much useless, as discussed above.

Third, in a great deal of the research where frequentist significance testing is applied, we are dealing with populations, not samples. In fact, in some of those analyses—conflict event data sets such as MID and ACLED are good examples—researchers go to a great deal of trouble making sure that they have all of the cases, and spend an inordinant amount of time tracking down the last 10% or so, suggesting that this comprehensiveness is considered a very important aspect of the overall research approach. In other instances, such as analyses of legislation or Supreme Court rulings, the population is clear and—with the Web—readily available. Short of major philosophical gymnastics—and again, these are clearly not how we really think—it becomes very difficult to figure out how one interprets the nominally sample-based significance test, as there is no sample. This, in turn, may also contribute to the tendency to interpret the p-value as if it were a measure, inferred from all of the available data, of the strength of causality, rather than a measure of the likelihood that results would be found in a sample from a population generated under the conditions of the null hypothesis.

Two open issues: which prediction and which Bayesianism

Having established where we *don't* want to be, the next step would be to establish where we want to go—which will be dealt with shortly—but first there are a couple of places where I've been a bit vague—prediction and our great savior, Bayesianism—and some further clarifications (the first will be a lot clearer than the second) are in order.

Predicting *what*?

While the primacy of prediction is clear in the logical positivists, I have simultaneously argued that the logical positivists worked in a deterministic world, whereas we do not. Prediction in a deterministic world can be very, very precise—consider for example the "Pioneer anomaly"

> When all known forces acting on the spacecraft are taken into consideration, a very small but unexplained force remains. It appears to cause a constant sunward acceleration of $(8.74 \pm 1.33) \times 10^{-10} m/s^2$ for both spacecraft.
> (http://en.wikipedia.org/wiki/Pioneer_anomaly; accessed 21 August 2010)

A measureable error on the order of 10^{-10} following 30 years of observation: that's determinism.

In the social sciences, we're not in that situation, nor will we ever be. Our models will always contain errors for at least the following reasons:

- Specification error: no model can contain all of the relevant variables;
- Measurement error: with very few exceptions,[36] variables will contain some measurement error, this presupposing there is even agreement on what the "correct" measurement is in an ideal setting;
- Free will: A rule-of-thumb from our rat-running colleagues in the behavioral sciences, "A genetically standardized experimental animal subjected to carefully controlled stimuli in a standardized laboratory setting, will do whatever it wants." This applies to humans as well and can result, for example, in "path dependent" behaviors;
- Quasi-random structural error: Complex and chaotic systems under at least some parameter combinations, as discussed below.

Since we can't depend on deterministic definitions of prediction, and merely "significantly different" is also not meaningful in a large set of situations where we would like to study political behavior, what is available? I think there are at least three possibilities, here.

First, we could use the norm that has emerged in meteorology—a natural science characterized by specification, measurement and structural error, but, unless we accept the Thor theory, not free will: The predicted probabilities and observed occurrence (assessed in random subsets of the data or out-of-sample) fall on a 45-degree line: that is, rainstorms predicted to occur with a 20% probability should be observed 20% of the time. This is the approach adopted by PITF (and to a limited extent, ICEWS): it requires models that produce probabilities but these are readily available, e.g. logit. King and Zeng (2001) provide an excellent illustration of this in correcting the flawed State Failures Project model.

Second, the predicted impact of changes in variables should be consistent with the distribution of change predicted by the model. This could be as simple as assessing the implications of the standard error of the estimate in a frequentist model—this will be normally distributed—or through a more complex posterior from a Bayesian estimation. Ideally, this can be assessed through natural experiments in out-of-sample tests, but in their absence, by using sub-samples in a population data set;

Third, in many classification models, the criterion becomes the ROC curve—again, this is the method now used in PITF analyses. In particular, this approach is very good at determining the degree to which a model does better than would be expected by chance (a test that astrology, for example, fails).

This list is probably not exhaustive, but the core point is that we don't have a workable probabilistic model of reasoning from either the logistical positivists or the frequentists. And we also know more than they did. Take, for example, the results of chaos theory which show that even a very simple non-linear deterministic system—you can get this in a

[36]There *are* exceptions: for example I doubt that the COW contiguity data has any error, unless there are typos. A data set defined, for example, by the word frequencies in the bills downloaded from the U.S. Library of Congress THOMAS database would—barring unlikely electronic glitches—be free of measurement error. But such situations are unusual.

difference equation with a single quadratic term—will exhibit seemingly random behavior for some parameter combinations. The logical positivists were largely unaware of chaos theory: while Poincaré had established the earliest results in the context of astronomical systems in the 1880s, and these had figured in theoretical work in probability theory in the first half of the 20th century, the appreciation of the extent to which these models applied in real-world systems did not occur until the works of Lorenz and Mandelbrot in the 1960s, and the topic was not widely studied until the 1980s. All of this work is well past the heyday of the logical positivists. While less mathematically cohesive, the same can be said for complexity theory, another development of the 1980s. The point here is that even if a social system were deterministic (per the logical positivists), the behaviors we observe will *appear* indistinguishable from a system that is stochastic.

WinBUGS Bayesianism or folk Bayesianism?

The Bayesian alternative solves numerous problems: it is logically coherent, and consequently it *can* provide the basis for a proper theory of inquiry, it cleanly solves the issue of the integration of theory and data, it corresponds to how most people actually think (which is no small advantage when one is trying to develop models of the behavior of thinking human beings), it solves the logical positivists' evidentiary paradox of yellow pencils being relevant to black crows, as well as the fact that researchers find some cases more interesting than others, and it provides a straightforward method of integrating informal *a priori* information with systematic data-based studies.

The downside to Bayesian approaches is mathematical and computational complexity. The latter now has purely technological fixes, though truth be told, 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.

There is a serious question, then, whether we need strict Bayesianism, or merely a less restrictive "folk Bayesianism" (McKeown 1999) that drops the most objectionable aspects of frequentism—the tyranny (and general irrelevance) of the significance test (whether knife-edge or p-valued) in the "let's pretend" world of the null-hypothesis, and the incompatibility of the sample-based assumptions of frequentism with the population-based analyses that characterize large parts of political science research. One can get past those constraints without necessarily jumping into all of the logical implications of Bayes Theorem.

Related to this issue—I think—is the task of developing modes of inference/interpretation for a much broader class of systematic patterns in data than can be found using frequentist or Bayesian analysis. We're already effectively doing this (though not admiting it) whenever we run an estimation on a population—for example, everything that has ever been done with Correlates of War data in any of its manifestations—because the frequentist interpretations are logically unsupportable (however popular) and, effectively, one has calculated descriptive rather than inferential statistics.[37] But, at present, these interpretations are *not* accepted in computational pattern recognition methods: for example the uncertainty over whether correspondence analysis is inferential or descriptive may be one of the reasons it is used

[37]Unless the inference is to an out-of-sample future you are predicting. But quantitative IR scholars don't like to predict, right?

infrequently on this side of the Atlantic. Yet many of these "data mining" methods[38] have much greater potential for predictive validity than the predominant frequentist approaches.

So where do we go from here?

This brief survey suggests to me that we have two closely inter-linked—but thus far largely disparate (and this may account for much of the motivation for this panel) —items to address. These developments need to occur concurrently though to date they seem to be occurring separately.

First, if frequentism is dead, or at the very least, a dead end, then bury it. The Bayesian alternative is sitting in plain sight, has been the focus of probably a majority of contemporary statistical research for at least twenty years, and at elite levels its superiority is already fully acknowledged.

The problem at the moment, however, is that we still teach statistics from the ABBA perspective: the Bayesian approach has won the heights but not the surrounding fields, and the fields are where most people, be they students or practitioners, are still living. We are approaching a situation where we've got a classical exoteric/esoteric knowledge distinction: we *teach* frequentism but, want to do state-of-the-art? Well, forget all that frequentist stuff, we're Bayesians. This is not healthy. Furthermore, the Bayesians themselves, obsessed with purely technological advances—that's you, Society for Political Methodology—have done little or nothing to change this situation: all of the major undergraduate "Scope and Methods" textbooks are frequentist.[39]

The pedagogical challenge here is, in fact, fairly severe. At present, a graduate student who has mastered KKV and, say, Maddala would be considered reasonably well prepared to handle almost all statistical methods encountered in mainstream political science. Modifying this by simple agglomeration—preparation to consist of Gerring for systematic qualitative methods, KKV except for the frequentist parts, Gill for Bayesian methods, Maddala for advanced linear models, and Theodoridis and Koutroumbas for computational pattern recognition, plus a specialized course or two in time series and hierarchical models (and by the way, that list still doesn't give you ANOVA or correspondence analysis)—is not realistic. Instead,

[38]This term is perjorative within political science but descriptive in many other data analysis communities; again see *The Economist* 2010. The difference is, in part, because commercial data-intensive methods—for example the Amazon and NetFlix recommendation algorithms—are consistently validated by out-of-sample testing. In a fixed sample, data-mining invokes "With six parameters I can fit an elephant and with seven I can make him wag his tail", whereas in out-of-sample evaluations, additional parameters are a potential *liability* due to the high risk of over-fitting, rather than an asset, which sharply constrains simple curve-fitting.

[39]See Turner and Thies 2009. My interpretation of their results is that the list is very concentrated: 5 author teams—Pollock, *The Essentials of Political Analysis* and/or *An SPSS/Stata Companion to Political Analysis* [27]; Johnson, Reynolds, and Mycoff, *Political Science Research Methods* [27]; Le Roy and Corbett, *Research Methods in Political Science* [11]; Carlson and Hyde, *Doing Empirical Political Research*[8]; Babbie, *The Basics of Social Research* or *The Practice of Social Research* [7]—account for nearly 50% of the 168 responses. Turner and Thies, however, conclude the opposite, noting "Though the survey reached a relatively small sample of courses, the results suggest that scope and methods instructors see a wide range of texts as suitable for their course. Indeed, though they could list more than one, 106 respondents produced a list of 71 distinct titles that they require for the scope and methods course. This textual variety suggests that instructors have found a multiplicity of suitable tools for conveying course content." (pg. 369)

a new approach that is selective—and quite possibly places more focus on fundamentals than on specific idiosyncratic techniques—is going to be needed.

Furthermore, the issue is much more complicated than just dropping the final two chapters of KKV: if you are going to do that, why not drop almost the whole of Maddala, which is frequentist? While we are doing a lot of nutty things with frequentist models—and believe me, the lack of a discernible response to King (1986) and Achen (2002) makes one sorely tempted to say "Sorry kids, we warned you, but you didn't pay attention and so we're going to have to take away the frequentist toys."—those techniques probably will continue to have considerable utility, at least for exploratory work on populations, and of course in survey and experimental situations for which they are generally appropriate. Given that much of the public face of political science, as a science, is survey research, we need to retain some elements of frequentism. I'm less convinced it will be relevant in policy settings where the interest tends towards causal factors: here we would probably be best off with a complete adoption of Bayesian interpretations, and some subset of Bayesian methods. But the larger point is that choices will need to be made, and the answers are not always obvious.

In our worst possible case, as Max Planck observed in his *Scientific Autobiography*, "a new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it." I'd like to think that progress can be made in my lifetime, but at the very least we need to start providing the materials for a new generation to work with.[40]

The second challenge is taking the philosophy-of-science ball from where it was dropped in mid-field by the logical positivists in the 1950s and moving on. We need a coherent theory—almost certainly based on Bayesian rather than frequentist principles—of stochastic inference concerning social behavior that is cleared of the detritus that accumulated by the attempts to achieve this solely through the modification of logical, deterministic theories.

In fact, a great deal of work *is* being done on these issues—for example by Gerring, Collier, Bennett, George, Geddes, Elman and others—though largely in the context of *qualitative* research. Those efforts have not, for the most part, incorporated the Bayesian perspectives, though for example many of the arguments I'm making here with respect to Bayesian interpretations of evidence that can be found in McKeown's (1999) decade-old critique of KKV. And to be perfectly clear, the Bayesian philosophical ball was dropped by the quantitative side, where philosophical development has been essentially absent for the better part of thirty years, and when it is attempted, it is often done quite badly—almost without fail, the worst papers at the Political Methodology Summer Meetings are "conceptual." I'm as guilty of this neglect of foundational theory as anyone—and hey, give me a break, we've had these seriously cool new toys appearing on our desks year after year for two decades now and it is really hard to go back and read Quine and Popper, to say nothing of Bacon and Descartes, under those circumstances. But yes, I'm guilty. And I am the first to acknowledge that the "qualitative" theorists are completely correct that there is no philosophical justification

[40]The new NSF-funded Online Portal of Social Science Education in Methodology—OPOSSEM—may eventually provide a venue for this: given the glacial pace with which publishers are willing to risk the commercial development of new methods texts for the political science market—instructors are still using Babbie for godsakes!—the likelihood of a publisher taking a financial risk on the development of a Bayesian curriculum, which we are very unlikely to get right the first, or even fifth, time, is very close to zero. But the open-source/open-access approach of OPOSSEM might succeed.

for dividing the discipline on the basis of level-of-measurement: we need a single scientific philosophy of inquiry that encompasses both quantitative and qualitative variables.[41]

But we've still got a ways to go, for example in the development of a coherent theory of evidence that rejects both the radical skepticism of Hume (as well as the incoherent skepticism of the postmodernists) but as well the equally constraining requirements of large-sample frequentism (e.g. as found in KKV). Humans in fact derive information incrementally from small samples (again, Bayesian inference deals with this in a straightforward manner), and they also regularly (in fact almost inevitably) derive information from situations where the number of potential explanatory variables is much larger than the number of observed cases (and where, often as not, the correlation of indicators is an asset, not a liability). To think otherwise is to be drawn into only a small sector of possible analytical modes—linear models and/or genuinely controlled randomized designs—and not all possible modes.

My sense, however, is that we may be closer to workable solutions than might first appear—it is darkest before the dawn and all that. For during the period when we have necessarily been fighting off an onslaught of zombie nihilists—*Brains...we must consume more brains...*—the philosophy of science has moved on in ways we are likely to find generally compatible (and would do well to incorporate into our pedagogy, rather than acting like Kuhn is the latest cool thing). For example, the political methodology enthusiasm for Bayesian approaches has parallels in mainstream philosophy of science; "scientific realism" has introduced a number of pragmatic elements which provide more flexibility than the strict logic of the Vienna Circle, and the social sciences are now considerably more developed as true sciences, and thus it is more likely that problems specific to social science will be addressed.

Why should we care?

I will conclude this essay with two observations: first, why should we care, and second, should we be concerned that this critique (and earlier, Achen's comparable points) is from someone approaching retirement?

What is the big deal about this "science" thing, anyway? If, channelling Mary Baker Eddy, the Transcendental Meditation movement wishes to assert that it can be scientifically validated, why not just let it? Here, I suppose, I am in much the same position as the Vienna Circle in their attacks on the unscientific claims of Freudian psychology and political Marxism, or the modern skeptical movement (Sagan, Asimov, Gardner, Randi, Kurtz and others) in popular culture against, for example, the claims of Uri Geller, Babbie's mentor Werner Erhard, and various UFO cults: A distinctly scientific approach dating from the 17th century got us to a world that I rather enjoy—and I include political science, notwithstanding its somewhat premature adoption of the title, in that category—and no, I'm sorry, but not just anything qualifies.

That said, I am a pluralist with respect to *knowledge* about the political world. The scientific method is only one of many methods for learning about that world, and in many instances, not the most efficient. Human political activity requires human understanding (and, by the way, a great deal of reasonably accurate prediction), and as work in evolutionary

[41] But no, we're not going to surrender the title "APSA Organized Section on Political Methodology." Or at least not until we surrender the "APSA" part.

psychology is showing in increasing detail as we compare human social cognition to that of other primates, the political structures we have created are due in no small part to our cognitive abilities to create them. Some of this behavior is subcognitive—we do things without fully realizing the reasons (Freud was at least on the right track on this one)—but, in the words of the sociologist William Whyte, we do a lot of things for the reasons we think we do them. Over the past ten years, for example, a great deal of systematic research has established the importance of fairness norms and the willingness of humans to "irrationally" exact revenge when those norms are violated. But Homer had a pretty good grasp of that regularity as well.[42]

So as long as someone can persuade me that they have a good grasp of the empirical world (and acknowledge that there exists an empirical world), some reasonably systematic way of trying to understand that world, and can convey this in a coherent manner that is likely to withstand the scrutiny of other experts in the field, I'll listen. Which is to say that I'll accept John *"Math is the enemy"* Mearsheimer into my world even if he won't accept me into his.

But that doesn't make all knowledge scientific nor, in a pluralistic worldview, would we want it that way: we distinguish, after all, between physicists and engineers, between biologists and medical practitioners. Furthermore, the scientific approach towards the study of politics can, in my experience, take us places we could not go otherwise, and do so in a transparent and replicable manner that transcends the effort and genius of any one individual. The Political Instability Task Force and the Integrated Conflict Early Warning System projects have achieved consistent, out-of-sample predictive accuracy far beyond that of human experts. These systems were not easy to construct, and thirty years ago, it was not possible to construct them, despite concerted efforts to do so. But through the application of systematic data collection and statistical techniques, they can be constructed now.

I would, in fact, argue that we would be better off to simply acknowledge that the "science" in "political science" was nothing more than a century-old steam-punk marketing ploy[43] and move, still in a spirit of pluralism, to a "small tent" definition of science that provides a clear and cohesive definition for what constitutes a scientific approach (which, to reiterate, will not divide on levels-of-measurement) but does not attempt to become so expansive as to include everyone (or even most people) teaching in departments labeled "Political Science." Interpretivist methods—which I fully acknowledge can provide useful insights into political behavior—are not scientific because they fail the test of having a systematic and replicable methodology. Rational choice theory, despite its use of mathematics[44] and its heuristic value in some situations such as the prisoners' dilemma and ultimatum games, in most instances fails miserably to cross the bar of predictive validity. A small tent

[42]The Greek poet, not Simpson. But come to think of it, Homer Simpson also qualifies.

[43]And the cachet of that label has probably diminished, and not just due to L. Ron Hubbard: the technological developments of the last quarter of the 20th century with the greatest impact on day-to-day life in the developed world—personal computing and computer networks—were almost exclusively the result of applied engineering and creative art (e.g. computer programming generally, and purely conceptual breakthroughs such as graphical interfaces, the internet, packet switching and the iPod)—rather than advances in basic science. In contrast, the late 19th century revolutions in electromagnetic communication, industrial chemistry and hybrid crops were driven by basic science.

[44]Sorry folks but just using mathematics isn't sufficient: so does astrology. And that is not trivial.

approach will in all likelihood sharpen both our research and our pedagogy, and differentiate the strengths of the genuinely scientific approaches from other approaches. The panels of the National Science Foundation, for example, already apply such a "small tent" criterion to applications for funding, though this is as often implicit as explicit, at least in my not-inconsiderable experience.[45]

Finally, as a registered old fart who in any reasonably prudent traditional culture would have long ago been strapped into a leaking canoe and sent out with the tide, should we be worried that this hybrid of a manifesto and diatribe—much like that of Achen (2002)—is being written by someone old, fat, happy, long-tenured, a Boomer, listening to *Jan & Dean* on `pandora.com`[46] and sounding more than a bit like "Damn young whippersnappers. . . can't even interpret a simple ANOVA can they?. . . piling all their silly papers in my in-box. . . get off my yard. . . "

Three thoughts on this issue. First, without question, I would be absolutely delighted if rather than Schrodt and Achen telling people to stop running kitchen-sink models, there was an uprising from the untenured (or, realistically, newly tenured) masses saying "We ain't going to run 12-variable linear models no more." To an extent, I suppose, we have seen this in the rejection of many of the prevailing quantitative approaches by some younger scholars, but in more than a few instances they have thrown the baby (and the dishpan) out with the bathwater.

Second, I simply see so much of this stuff as a reviewer and discussant, and over the years I've seen so many methodological fads come and go. Really, a mind is indeed a terrible thing to waste, particularly on running the 3,001st analysis of the Oneal-Russett data set (or anything remotely resembling the Oneal-Russett data set). The parallels with Scholasticism are also very scary: from 1650 onwards European society experienced tremendous political, cultural and technological progress, and yet until the late 19th century the attitude within the universities was "Party like it's 1274 A.D.!" I am completely confident of the ability of the academic world to resist change, but I don't necessarily think this is a good thing.

Finally, there is a old aphorism—it may be due to Arthur C. Clarke—that when a senior member of a profession says something is impossible, it probably isn't; when a senior member of a profession says something is possible, it probably is. We're in a bit of a mess at the moment, but the way out is not only fairly clear, but we are already well along on that way. So I'm saying that a resolution is possible.

[45]There's an interesting parallel here with...okay, indulge me for a moment...the Council of Nicæa in 325. Despite its eventual theological importance—the Nicæan Creed and its derivatives are recited to this day by Christian congregations around the world—Nicæa was called by a political authority to solve a political problem: Constantine having decided to provide monetary support (and, by the way, hold the lions. . .) to Christian institutions needed to figure out exactly who did and did not qualify as a Christian, and with the fiscal clout of the Roman Empire in play, just calling yourself "Christian" was not sufficient. The NSF is performing a similar role. Not that I'm suggesting we write a creedal affirmation.

[46]But I'm also listening to *Arcade Fire*. Honest!

References

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

[2] Achen, Christopher. 2002. Toward a New Political Methodology: Microfoundations and ART. *Annual Review of Political Science* 5: 423-450

[3] Beck, Nathaniel, Jonathan N. Katz and Richard Tucker. 1998. Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable. *American Journal of Political Science* 42:1260-1288.

[4] Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning.* Springer.

[5] Brady, Henry and David Collier, eds. 2004. *Rethinking Social Inquiry.* Lanham, MD: Rowman and Littlefield.

[6] Davies, J. L. and T. R. Gurr. (Eds.). (1998). *Preventive Measures: Building Risk Assessment and Crisis Early Warning.* Lanham, MD: Rowman and Littlefield. (This has chapters on most of the major forecasting projects developed in the 1990s.)

[7] Duda, Richard O., Peter E. Hart and David G. Stork. 2001. *Pattern Classification*, 2nd ed. Wiley.

[8] *The Economist.* 2010. Data, Data Everywhere: A Special Report on Managing Information. *The Economist.* 27 February 2010.

[9] Freedman, David A. 2005. *Statistical Models: Theory and Practice.* Cambridge University Press (2005)

[10] Freedman , David A., David Collier, Jasjeet Sekhon and Philip B. Stark, eds. 2009. *Statistical Models and Causal Inference: A Dialogue with the Social Sciences.* Cambridge: Cambridge University Press.

[11] Gill, Jeff. 1999. The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly* 52:3, 647-674.

[12] Gill, Jeff. 2003. *Bayesian methods : a social and behavioral sciences approach.* Boca Raton, FL : Chapman and Hall.

[13] Goldstone, Jack A., 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.

[14] Hempel, Carl G. and Paul Oppenheim, "Studies in the Logic of Explanation." *Philosophy of Science* 15,2: 135-175.

[15] Hempel, Carl G. 2001."Explanation and Prediction by Covering Laws." Chapter 5 in Carl G. Hempel and James H. Fetzer (ed) . *The philosophy of Carl G. Hempel : studies in science, explanation, and rationality.* Oxford: Oxford University Press.

[16] Huff, Darrell. 1954. *How to lie with statistics.* New York: Norton.

[17] King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry.* Princeton University Press.

[18] King, Gary and Langche Zeng. 2001. Improving Forecasts of State Failure. *World Politics* 53(4): 623-658.

[19] King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30, 3: 666-87.

[20] Lazarfeld, Paul F. 1937. Some Remarks on Typological Procedures in Social Research. *Zietschrift Fuer Sozialforschung* 6: 119-39

[21] Markovsky, Barry and Fales, Evan. 1997. Evaluating Heterodox Theories. *Social Forces* 76, 2:511-25.

[22] McKeown, Timothy. 1999. Case Studies and the Statistical Worldview: Review of King, Keohane, and Verba's *Designing Social Inquiry: Scientific Inference in Qualitative Research. International Organization* 53, 1: 161-190.

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

[24] O'Brien, Sean. 2010. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review* 12,1:.87-104

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

[26] Orme-Johnson, D.W., Alexander, C.N., Davies, J.L., Chandler, H.M., and Larimore, W.E. 1988. International peace project in the Middle East: The effects of the Maharishi Technology of the Unified Field. *Journal of Conflict Resolution* 32: 776-812.

[27] Quine, Willard Van Orman. 1951, "Two Dogmas of Empiricism." *The Philosophical Review* 60: 20-43.

[28] Political Analysis. 2004. Special Issue of Political Analysis on Bayesian Methods. *Political Analysis* 12:4

[29] Richardson, Lewis F. 1960. *Statistics of Deadly Quarrels.* Chicago: Quadrangle.

[30] Schrodt, Philip A. 2006.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.

[31] Schrodt, Philip A. 2006. Beyond the Linear Frequentist Orthodoxy. *Political Analysis* 14,3:335-339.

[32] Schrodt, Philip A. 2009. Reflections on the State of Political Methodology. *The Political Methodologist* 17,1:2-4.

[33] Sokal, Alan D. 1998. *Fashionable Nonsense: Postmodern Intellectuals' Abuse of Science.* Picador.

[34] Tetlock, Philip E. 2005. *Expert Political Judgement.* Princeton: Princeton University Press.

[35] Theodoridis, Sergios and Konstantinos Koutroumbas. 2009. *Pattern Recognition*, 4th ed. Springer.

[36] Turner, Charles C. and Cameron G. Thies. 2009. What We Mean by Scope and Methods: A Survey of Undergraduate Scope and Methods Courses *PS* April 2009: 367-373.

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

[38] 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*:47,5.