Puzzles, Proverbs, and Omega Matrices: The Scientific and Social Significance of Empirical Implications of Theoretical Models (EITM)

By Jim Granato and Frank Scioli

> "There is considerable intellectual ferment among political scientists today owing to the fact that the traditional methods of their discipline seem to have wound up in a cul-de-sac. These traditional methods—i.e., history writing, the description of institutions, and legal analysis—have been thoroughly exploited in the last two generations and now it seems to many (including myself) that they can produce only wisdom and neither science nor knowledge. And while wisdom is certainly useful in the affairs of men, such a result is a failure to live up to the promise in the name political *science*." —William H. Riker[1]

What has changed since William Riker's statement was written?[2] No doubt Political Science has made considerable headway in going beyond traditional methods. For example, social choice theory, formal models of party systems, and statistical methods for analyzing electoral and roll-call data have made it possible to make scientific progress in the study of democratic institutions and party systems.

Yet, before we pop the champagne cork and exclaim to the world "we have arrived," there are some common and current research practices that, if continued, could delay, or worse, derail the momentum generated over the past 40 years. What is at stake is fundamental, as there is a risk that in the future the political science literature may come to be characterized as a proliferation of noncumulative studies.

The term noncumulation means that if certain research and teaching practices continue and become dominant, we will find ourselves with a limited foundational basis for evaluating our models, unsure of their strength and too often unable to know where they went wrong. The problem can only be exacerbated if we proceed apace in what amounts to a research assembly line where quantity substitutes for scientific quality. Repetition may be good for many things but it does not serve as a substitute for scientific progress.

None of this should be surprising. The recent *Report of the APSA Ad Hoc Committee on the National Science Foundation* found that political science had been characterized as, "not very exciting, not on the cutting edge of the research enterprise, and in certain quarters as journalistic and reformist."[3] We disagree with this statement and believe there has been considerable improvement in political science in the past 40 years through the use of formal models, case studies, and applied statistical modeling.

However, there is a danger of advancing these methods without advancing our scientific understanding of politics. Each approach has individual strengths and weaknesses:

- Formal models force clarity about assumptions and concepts; they ensure logical consistency, and they describe the underlying mechanisms that lead to outcomes.[4] They also can lead to surprising results, such as the free rider problem or the power of the median voter, which have spawned substantial literatures. However, formal models can fail to incorporate empirical findings in order to provide a more accurate depiction of the specified relations. The models may be elegant, but too often they ignore, or even throw out, useful information. This results in modeling efforts that yield inaccurate predictions or do not fit findings. In fact, data may contradict not just a model's results but also its foundational assumptions.

Jim Granato is Visiting Scientist for Political Science in the Social, Behavioral and Economic Sciences Directorate at the National Science Foundation (jgranato@nsf.gov). Frank Scioli is Program Director for Political Science in the Social, Behavioral and Economic Sciences Directorate at the National Science Foundation (fscioli@nsf.gov). *The authors thank Chris Achen, John Aldrich, Bill Bernhard, Norman Bradburn, Brian Humes, Mark Jones, Richard Lempert, Phil Shively, Joan Sieber, Duncan Snidal, and Paul Wahlbeck for their comments on earlier versions of this paper. They also thank Henry Brady and Jennifer Hochschild for their comments and assistance through this process. We dedicate this paper to the memory and scientific accomplishments of EITM Workshop participant Richard McKelvey.*

- Case studies provide detailed information about the steps by which events occur and allow researchers to identify mechanisms that can produce such phenomena as groupthink, authoritarian regimes, revolutions, and ethnic conflict. Case studies also enable researchers to discover enough about countries to distinguish idiosyncratic from general causes, to identify interactive and connected causes, and to understand how people's interpretations of events—the meaning that they have for people—affect their actions. But case studies sometimes focus too much on the idiosyncratic details of rare and influential events. They may miss the opportunity to inform a more general theory. In some instances the result amounts to theorizing by proverb: that is, site-specific theories expressed as causal theories.[5]
- Applied statistical models can provide generalizations and rule out alternative explanations through multivariate analysis. Researchers are forced to conceptualize putative causes so that they can be reliably measured. Models can distinguish between causes and effects, allow for reciprocal causation, and estimate the relative size of effects. Yet many of the best methodologists wonder if we have reached the point of diminishing marginal returns with statistical analysis. The variables in regressions are sometimes poor reflections of theoretical concepts. Empirical models often seem more like data mined "garbage-can regressions and garbage-can likelihoods" because of their lack of causal motivation and theoretical specificity.[6] Indeed, model shortcomings are typically treated as statistical problems requiring statistical patches, instead of a more careful specification of the mechanism behind the model. The distance between theory and test can only grow with this mindset.

What can be done? One way to address these problems is to change standard research practices and enhance training opportunities so that formal and empirical analyses (applied statistical *and* case study analyses) are linked. Large-N analysis can test a formal model through statistical analysis, and small-n case studies can also test a model by seeing if the mechanism postulated by the model really exists.

Against this background, the NSF's Political Science Program recently developed an initiative to link formal and empirical analysis that is called Empirical Implications of Theoretical Models (EITM).[7] The objective of EITM is to encourage political scientists to build formal models that are connected to an empirical test. As scholars merge formal and empirical analysis, we think they lay the groundwork for (social) scientific cumulation. Why? By thinking about the empirical implications of theoretical models, scholars develop clear-cut empirical tests of the models. This symbiosis means that concepts must be clarified, and causal linkages must be specified. Theories must meet the challenge of these tests, and empirical work must be linked to a theory. Theories and concepts that fail are discarded. Empirical work reveals the range and the limitations of theory. Useful generalizations are produced, and political science becomes worthy of its name.[8]

We are not there yet. In the rest of this essay we describe the far too lax practices in the way political scientists usually endeavor to put their concepts into operation and to establish causation. This necessitates discussing some of the scientific and social consequences of several common research practices in the discipline. It also requires an in-depth explanation of EITM, and the provision of examples in the political and social sciences that demonstrate EITM's capacity for improving our ability to establish valid causal linkages. Finally, we discuss EITM-inspired improvements in technical-analytical competence.

The Scientific Consequences of Current Research Practices: A Role for EITM

At the most basic level, formal modeling assists in the "construction of valid arguments such that the fact or facts to be explained can be derived from the premises that constitute the explanation."[9] An important virtue of formal modeling is that it often yields surprising implications that would not have been considered had they not emerged from formal analysis. Conversely, if practiced correctly, applied statistical and case study analysis shows researchers where a model went wrong and leaves open the possibility that a more accurate model can be constructed.

While the emphasis of EITM is on the linkage between formal and empirical analysis and is a natural fit with a quantitative approach to research and teaching, we think it is a mistake to consider EITM as an exclusively "quantitative" or mathematical enterprise. Three related points can clarify how EITM fits into broad scientific inquiry:

- First, we believe that qualitative analysis (i.e., small-n case studies and the like) and quantitative analysis contribute to cumulative knowledge when they are thought of and used as mutually reinforcing methods. EITM is intended, in part, to encourage and accelerate shared standards and multiple method approaches.
- Second, models can be mathematical, but they do not have to be mathematical to be useful. All that is required is careful logical arguments that identify concepts and causal linkages. In addition, models need not be tested with statistical analysis. Case studies can be used to validate the concepts used in the model or to search for causal linkages predicted by it.[10] Maurice Allais, a scholar known for a mathematical approach to his research questions, makes a related point:

  > Genuine progress never consists in a purely formal exposition, but always in the discovery of the guiding ideas which underlie any proof. It is these basic ideas which must be explicitly stated and discussed. Mathematics cannot be an end in itself. It can be and should only be a means.[11]

- Third, because we believe in uniting qualitative and quantitative approaches, the question arises of how this unification and ultimate relation to EITM can come about. We think the answer is that qualitative analysis serves as a complement to quantitative analysis on both the formal[12] and applied statistical levels.[13] We also see work

underway to establish this complementarity on both the theoretical[14] and empirical sides.[15]

In an ideal world, where there is unification in approach, political science research should have the following components: (1) theory (informed by case study, field work, or a "puzzle"); (2) a model identifying causal linkages; (3) deductions and hypotheses; (4) measurement and research design; and (5) data collection and analysis. What we find is that because they are generally treated by scholars as distinct, separable approaches, the three most common current research practices—formal modeling, case study analysis, and applied statistical modeling—deviate from this ideal. They therefore limit the possibilities for substantial enhancement of knowledge.[16]

Further, as we noted earlier, three common shortcomings may occur when researchers limit themselves to the strengths and weaknesses of a single methodological approach. For formal modelers, this manifests itself in not respecting the facts; for qualitative researchers, this can result in theorizing by proverb; and for researchers who rely exclusively on applied statistics, we find data mining, garbage-can regressions, and statistical patches (i.e., omega matrices).

Respecting the facts

The assumptions on which some formal modeling rests are often so at variance with empirical reality that model results are dismissed out of hand by those familiar with the facts. The problem is not just unreal assumptions, for one way to build helpful models is to begin with stylized and perhaps overly simple assumptions, test the model's predictions, and then modify the assumptions consistent with a progressively more accurate model of reality. Yet these follow-up steps are too often not taken or left unfinished, with the result being a model that does little to enhance understanding or to advance the discipline.

One justification for "theories of unreality" is that realistic models are often so complex as to be of limited value. There is merit to this defense. An important function of formal modeling is to assist in identifying crucial quantitative and qualitative effects from those that are of minimal importance. However, the drive for simplicity can be taken too far. This conflict between realism and analytical tractability is not new or only a problem in the discipline of political science. Economics is instructive in this regard. In the late 1960s and early 1970s there was a revolution in macroeconomic research, which put great emphasis on the microfoundations of macroeconomic outcomes. Yet, as George Akerlof recently noted, "[T]he behavioral assumptions were so primitive that the model faced extreme difficulty in accounting for at least six macroeconomic phenomena. In some cases logical consistency with key assumptions of the new classical model led to outright denials of the phenomena in question; in other cases, the explanations offered were merely tortuous."[17]

The use of simplifying assumptions is in principle a virtue and remains so when such simplifications do no harm to overall predictive accuracy. However, this does not mean that formal modeling should proceed without regard to glaring factual contradictions in its foundational or situation-specific assumptions. Rather, formal modelers must be especially careful to make sure that they test their models in situations that go beyond the circumstances that suggested the models, for it is there that simplifying assumptions are likely to lead to difficulties.

In this situation, the role for EITM is to enhance the formal analysis by adding the necessary empirical results that lead to more accurate representations. When this approach is adhered to, researchers can ensure that simplifications do not compromise scientific rigor by leading to models that only predict the phenomena that led to their development in the first place.

Theorizing by proverb

For case study analysis, Riker's admonition still applies when practice degenerates into theorizing by proverb. While such studies can be richly illuminating, it is sometime hard to know where empirical tests leave off and the researcher's perspective or biases begin. Nor can one know what salient facts might be left out, or whether apparently strong findings represent idiosyncrasies of time and place rather than powerful general tendencies.

When a researcher goes from an empirical puzzle to a theory and then to hypothesis testing using the same observations, there is no guarantee that the observed relations support the theory. Indeed, they cannot be said to support the theory if it was inductively derived from the case studied. Yet, case studies often produce very useful theoretical propositions that perhaps reflect V.O. Key's observation of 50 years ago that "uniformities of behavior, at least in the aggregate, do turn up with astonishing regularity."[18]

Consistent with Key's assertions, we think a researcher's task is to take the reliably measured aspects of behavior and then attempt to provide valid generalizations. If a researcher can point to a set of conditions or circumstances where a variable X occurs and a variable Y changes in response, then it is possible to construct theories that serve vital purposes. For policy makers, who rely on theories and findings to take action, it is not just a question of seeing a specific consequence to their decision and action; rather, it is also about knowing the

For policy makers, who rely on theories and findings to take action, it is not just a question of seeing a specific consequence to their decision and action; rather, it is also about knowing the circumstances in which a policy choice is likely to produce a desired outcome, as well as when it is wise to take an action.

circumstances in which a policy choice is likely to produce a desired outcome, as well as when it is wise to take an action.

In non-systematic case study analysis there is a distinct danger that in generalizing to a specific causal relation, the researcher will create proverbs rather than theories. Herbert Simon warned long ago that proverbs, for all their uses as rhetorical devices, fail to foster scientific progress:

> It is not that the propositions expressed by the proverbs are insufficient; it is rather that they prove too much. A scientific theory should tell us what is true but what also is false. . . . For almost every principle one can find an equally plausible acceptable contradictory principle. In this situation there is nothing in the theory to indicate which is the proper one to apply.[19]

The danger exists that unless carefully and systematically undertaken, the case study method lacks sufficient power to separate out the many plausible rival explanations. Many factors may cause variation in a dependent variable, not just the plausible mechanism for which researchers might effectively argue. When we question explanations rather than accept them because of surface plausibility, we are able to find out which of the many potential factors matter most and how causes are conditional by context.

There are other problems as well.[20] When not executed correctly, case study analysis can degenerate into a process of conceptual proliferation and redefinition. While classification and definition constitute a foundation for general rules of relationships and sequential patterns, that is not the usual outcome.

William Riker provides guidance on this matter. In his book, *Liberalism Against Populism,* he examines the concept of "democracy." Rather than settle on a single definition, he asserts, "we cannot go to a unique authoritative source for a definition."[21] As a substitute, he examines commonalities in documents and lists:

> [He seeks] the properties found in these documents . . . [that are] elements of the democratic method . . . [and that] are means to render voting practically effective and politically significant, and all the elements of the democratic ideal [that] are moral extensions and elaborations of the features of the method that make voting work.[22]

To take Riker's approach one step further, we are arguing (using EITM) that when a researcher formalizes a model and then collects the necessary data to link the model with empirical outcomes, s/he has taken a crucial step toward identifying causal linkages. That, after all, is the ultimate goal of scientific endeavors.

Current and common applied statistical practices

When applying statistics, analysis too often turns into garbage-can models and jerry-built statistical contraptions with little if any relation to the theory. In *A Primer of Statistics for Political Scientists,* Key makes the case for the utility of quantitative analysis:

> A general point of view of the book is that quantitative procedures may be best regarded as particular techniques by which more general methods of reasoning may be applied to the data of politics. Hence, the discussion may be suggestive to students concerned about systematic political analysis regardless of whether they need to learn the intricacies of quantitative technique. A virtue of the statistical approach is that it brings explicitly and nakedly to attention general questions of analytical method.[23]

However, in the current research environment where data mining, garbage-can specifications, and statistical patches dominate, the ability to harness the attributes noted above—particularly generalizability—is compromised.

To be more specific, the following ratio is the subject of much attention by applied statistical analysts because it is the basis for which "theories" survive or perish:

$$\frac{b}{\text{s.e.}(b)}$$

This ratio is commonly referred to as a "t-statistic." It is the "truth" that most applied statistical analysts are concerned with, and it can be confounded by influences that shift the numerator (b) in unforeseen ways. The denominator, the standard error [s.e.(b)], also is susceptible to numerous forces that can make it artificially large or small. In either case, avoiding false rejection of the null hypothesis (Type I error) or false acceptance of the null hypothesis (Type II error) is imperative. While the concern with Type I and Type II errors should be of prime importance, that unfortunately is not usually the case. Instead, the focus is on the size of the t-statistic and whether one can get "significant" results.

The first tendency in trying to achieve "significant" results is the practice of data mining. Some political scientists put data into a statistical program with minimal theory and run regression after regression until they get either statistically significant coefficients or coefficients that they like. This search is not random and can wither away the strength of causal claims.

A second practice is that many studies degenerate into garbage-can regression or garbage-can likelihood renditions. By a garbage-can regression or likelihood we mean a practice whereby a researcher includes, in a haphazard fashion, a plethora of independent variables into a statistical package and gets significant results somewhere. But a link with a formal model could help in distinguishing the variables and relations that matter most from those that are ancillary and, probably, statistical artifacts. More often than not there is little or no attention paid to the numerous potential confounding factors that could corrupt statistical inferences.

The first and second practices lead to the third—statistical patching (i.e., the use of weighting procedures to adjust the standard errors [s.e.(b)] in the t-statistic ratio above). Data mining and garbage-can approaches virtually are guaranteed to break down statistically. The question is what to do when these failures occur. We think the answer to this question involves returning to the drawing board and developing new and more accurate specifications (with the aid of a formal model). However, this is not the usual practice.

Consider again the t-statistic above and the incentive to achieve significant results. One can do this by using statistical patches that have the potential to *deflate* the standard error and *inflate* the t-statistic, which, of course, increases the chance for statistical significance. Unfortunately, with the advances in computing power and the simplification of statistical software packages, this practice is only a click on the "enter" key away.

There are elaborate ways of using error-weighting techniques to "correct" model misspecifications or to use other statistical patches that substitute for a new specification. For example, in almost any intermediate econometrics textbook[24] one finds a section that has the Greek symbol Omega (Ω). This symbol is representative of the procedure whereby a researcher weights the data that are arrayed (in matrix form) so that the statistical errors, and ultimately the standard error noted above, are sometimes reduced in size and the t-statistic then may become significant.

In principle, there is nothing wrong with knowing the Omega matrix for a particular statistical model. The trouble comes in how one uses it. Consider that Omega matrices remove or filter residual behavior. The standard error(s) produced by an Omega matrix should only serve as a check on whether inferences have been confounded to such an extent that a Type I or Type II error has been committed.

Far too often, however, researchers treat the Omega weights (or alternative statistical patches) as the result of a true model. This attitude hampers scientific progress because it uses a model's mistakes to obscure flaws. Achen points out the scientific problem with these empirical practices:

> Why should anyone believe such research? Actually, and contrary to what disciplinary outsiders sometimes think, we don't. Article by article, we profess belief and sometimes manage to convince ourselves; but in truth, not much of real theoretical power cumulates after years of work. Result: we have no first principles to teach undergraduates and no agreed foundation from which to talk to policymakers about voter turnout, campaign finance, or how well democratic representation is working. We don't know and they know we don't know. We claim to be studying our data. But real data analysis has a character incompatible with most of our current practice.[25]

The recent controversy over the United States' presidential election models illustrates the problem with current empirical practices. Despite the fact that Tse-min Lin has demonstrated the overall lack of robustness of the most widely used models, there has been little movement in reformulating key features of these models.[26] Moreover, although such models rely on economic factors, the lack of intellectual introspection leaves the revolutionary advances in the past 40 years in empirically supported formal macroeconomic theory out of this analysis, despite their potential relevance to the issues political scientists confront and the known scientific quality of the new economic theories.

Case study analysis can also illuminate weaknesses in current applied statistical practice as they pertain to the 2000 presidential election. Henry Brady notes flaws in John R. Lott's statistical analysis of Florida voter turnout for the 2000 presidential election.[27] Lott's statistical findings indicate that the early call effect by the major news organizations (most of the Florida Panhandle is in the Central Time Zone, while the majority of the state is in the Eastern Time Zone) resulted in 10,000 fewer Republican votes than had been the case in the three prior elections. However, Brady examines Lott's results using an in-depth case study approach that "draws upon multiple sources of information, utilizing inferences based on common sense."[28]

Using back-of-the-envelope "case-study" kinds of calculations, including census data, uniformity of voting during the last hour (when the results were declared), media exposure on who would have heard the early call, and other factors, Brady's analysis shows that the figure is more likely to be between 28 and 56 votes lost for Republicans. This is an example where "causal-process observations demonstrate that it was highly implausible for the media effect suggested by Lott's analysis to have occurred."[29]

If one were to summarize the problem here, one would conclude that the intellectual drift from the virtues of empirical practices means that statistical technique has come to dominate the practices used to help identify causal linkages. But statistical technique alone cannot test generalizations of observed political behavior. Once again, the solution is to find ways to link statistical techniques with formal theory:

> Traditionally we have tried to do both with informal assumptions about the right list of control variables, linearity assumptions, distributional assumptions, and a host of other assumptions, followed by a significance test on a coefficient. But since all the assumptions are somewhat doubtful and largely untested, so are the estimators and the conclusions. The depressing consequence is that at present we have very little useful empirical work with which to guide formal theory. The behavioral work too often ignores formal theory. That might not be so bad if it did its job well. But it produces few reliable empirical generalizations because its tests are rarely sharp or persuasive. Thus, empirical findings accumulate but do not cumulate.[30]

The Societal Consequences of Current Research Practices

We think policy makers want to examine the best scientific evidence relating to the issue before them. What will they find when they turn to their shelves to look for research findings that are reliable and valid and provide identifiable predictions?

There are real world examples where the failure to integrate formal and empirical analysis can lead to predictive failure. For instance, Milton Friedman describes an experience he had while working for Columbia University's Statistical Research Group during World War II. Friedman "was to serve as a statistical consultant to a number of projects to develop an improved alloy for use in airplane turbo-chargers and as a lining for jet engines."[31] One task was to determine the amount of time it took for a blade made of an alloy to fracture. At the most basic level, Friedman relied on data from a variety of lab experiments to assist him in addressing this problem. He then used the data to estimate a single equation linear regression. This linear regression equation expressed time to fracture as a function of stress, temperature, and variables describing the composition of the alloy. Standard statistical indicators suggested his approach was valid. The analysis predicted that the blade would rupture in "several hundred hours." Yet the results of actual laboratory tests indicated that a rupture occurred in "something like 1–4 hours."[32] Because of the lab test results—and not the linear regression—the alloy was discarded.

Since Friedman relied primarily on an empirical (applied statistical) approach, he probably missed some important interaction effects or even some important variables that could have

predicted the rupture. This "surprising" failure might have been avoided if Friedman had used a statistical technique accompanied by a fully explicated formal model or multiple sources of information to guide his analysis.

Obviously, engineers should make sure their models of a particular structure (such as a bridge) are fully informed by theoretical understanding. Collapsing structures that are based on unsound scientific results potentially hurt hundreds. Is it not also obvious that weak scientific research findings presented by political and social scientists may do great damage to society (where not hundreds but thousands may be harmed) if policy makers use flawed findings as the basis for policy formulation?

EITM offers, through the integration of a formal model and empirical test, the specification of the conditions under which empirical possibilities occur. Consider the difference between knowing that cold weather can make water freeze and knowing the conditions of time and temperature it takes for water to freeze. Both forms of knowledge are correct, but there is considerable difference in precision and explanation.

Science, Society, and EITM-Type Research: Examples

We have argued that the scientific and social consequences of current research practices have led to the need for EITM. There are instances, in fact, in which the combination of formal with empirical analysis has resulted in scientific research with societal implications. The following examples are instructive in this regard.

Growth in partisan identification

Party identification research has clearly occupied an important place in political science.[33] Would one expect to see large changes in the stability of voting behavior or in the degree of party identification over the years that elections have been held in a new democracy? Is it possible to link the findings concerning party affiliation in Europe and the United States and generalize to an emerging democracy, or do cultural differences between the new democracy and these other countries result in different outcomes with respect to the intergenerational transmission of political attitudes?

We know that within advanced industrial democracies, there is generally not a high level of political and social awareness or political and social participation. The primary accountability mechanism—voting, for example—is usually limited to well-educated and prosperous segments of the society. Moreover, the typical voter tends to be ill-informed about political, social, and economic issues and about candidates' positions on these issues. Few voters take part in politics in ways other than voting, and their voting behavior seems mainly determined by group loyalties, such as social class, religion, ethnic affiliation, and above all, party identification.

How did we learn this? During the 1940s and 1950s the development and application of new social research techniques such as survey research (and subsequent computer analysis) provided new and sometimes startling insights into political behavior. When micro-analytic research methods began to be used in political science, it became possible to study the dynamics of individual voting behavior.

The rational actor model of representative democracy—the notion that citizens make rational decisions, which are the basis of their voting behavior—was challenged by what was discovered. Many paradoxes emerged. For one, political attitudes and voting behavior are determined, in part, by the social milieu into which one is born and the context in which one is raised. Although there are good historical reasons why certain religious, ethnic, or geographical groups vote as they do, in succeeding generations those political affiliations become fixed and are transmitted from one generation to another. They cease to be the result of rational analysis or current events.

Realizing this, political scientists began to study voting behavior more in relation to political socialization and less in relation to issue conflicts. In 1969, in his seminal article, "Of Time and Partisan Stability," Philip Converse advanced the theory that strength of party identification (and voting behavior) is primarily a function of intergenerational transmission plus the number of times one had voted in free elections.

Converse and Georges Dupeux had found in their earlier comparative study of France and the United States that 75 percent of Americans identified with a political party while only 45 percent of French did so.[34] "Other studies had shown high levels of party identification in Great Britain and Norway, and lower levels of party identification in Germany and Italy."[35] Converse and Dupeux further discovered that the difference in the percentage of party identifiers in France and the United States was explained in large part by the fact that more Americans than French knew the partisan identification of their father. For both countries, when citizens knew their father's party identification, about 80 percent of them identified with a party. Otherwise only about 50 percent did so.

In his subsequent research, Converse presented a theoretical framework in which he assumed that few of the fathers (in France) identified with a party. Proceeding with that conjecture, he reasoned that if the 50 percent of French voters who did not know their father's party identification became party identifiers, then the later generations would be more like Americans in that their partisan identification rates would be significantly higher. In short, France would become more like the United States, Great Britain, and Norway in party identification in ensuing generations and less like Germany, Italy, and earlier generations in France. To support this conjecture Converse made use of a mathematical technique: the Markov chain (see the Appendix).[36]

With the assistance of the Markov chain model, the theory has the following dynamic: In the first election in a country, almost no one would identify with a party. By the time the second-generation voters were of voting age, 50 percent of them would express identification. Then, gradually, party identification would rise to a stable level of about 72 percent. Converse assumed that the relatively low level of party identification in France resulted from the fact that the vote was extended later there than in the United States, and that French women were not given the vote until 1945.

Converse then noted that as individual voters become older, they identify more strongly with one party. This is not strictly a result of age, but of how long the person is exposed to a party by being able to vote for it. Combining these two findings, he came up with the theory that the strength of a voter's party identification can be predicted from two factors: The number of years the person was eligible to vote, and the likelihood that the individual's father had identified with a party.[37] Furthermore, several predictions can be derived from Converse's theory:

- When elections are first held in a country, little party identification is observed. (The pattern found in mature democracies—that older people have stronger party identification than younger people do—is not observed in the early days of a democracy.)
- If elections are interrupted, then a country's level of party identification declines.
- "If the transition rates for all countries are roughly the same as for France and the United States, then party identification levels in all electoral democracies should converge over a few generations toward a single value of about 72 percent."[38]

To summarize, Converse started with a fundamental puzzle that had important societal consequences, and then analyzed data. He then used a mathematical apparatus to generalize the case beyond the two-country comparison. At its most rudimentary level, he combined all the fundamental components of EITM-like research to enhance knowledge. In the ensuing years, a new generation of research has built on this beginning to provide more elaborate formal explanations about how individuals learn about political parties and events and adjust their partisan attitudes.[39]

The Phillips curve

Beginning in the early 1960s, macroeconomic policy makers (or their advisors) constructed statistical models to determine the effects of monetary and fiscal initiatives on unemployment and output. The scientific basis for this emphasis on policy "fine tuning" centered on the research of A. W. Phillips. Phillips showed empirically that there was an inverse relationship between nominal wages and unemployment: higher unemployment was associated with lower wages, while higher wages were associated with lower unemployment. This relation was extended to incorporate a trade-off between inflation and unemployment. From the late 1950s to the late 1960s most economists assumed that there was a stable trade-off between unemployment (or output) and inflation. In fact, this stable relationship could be graphically demonstrated on what is now called the Phillips curve.[40]

This assumption of a stable relationship had powerful appeal to policymakers. One could observe the corresponding rates of inflation and unemployment on the Phillips curve and formulate an appropriate monetary and fiscal policy to stimulate or contract the economy. The United States' experience with low unemployment and inflation in the early to mid-1960s suggested this was sound policy.

This optimism was to be short lived. Policy analysts noted by 1970 that their conditional forecasts were incorrect. The negative empirical relation between inflation and unemployment (or output) that Phillips found, and that was implicit in the fine-tuning policies, disintegrated in the 1970s. The prevailing conditions of high inflation *and* high unemployment, which came to be known as stagflation, defied the characterization offered by Phillips.

The policy failure was preordained, in large part by the failure to reconcile the empirical relations with standard models in economics. In the latter part of the 1960s, Friedman and Edmund Phelps, using both formal and non-formal theoretical arguments, demonstrated that the underlying Phillips curve assumptions were inconsistent with basic economic theory.[41]

In particular, Friedman and Phelps both emphasized that the Phillips curve is inaccurate when the public's expectations of inflation are taken into consideration. They argued that a stimulative policy could lower unemployment for a brief time if workers set their wage demands too low. This occurs if workers underestimate future inflation, but Friedman and Phelps reasoned that workers could not be fooled for long. They would eventually correct this mistake. During the transition to correcting the inflation expectation error, however, unemployment falls because wages have not kept pace with inflation, and employers can better afford the cheaper labor during this period.

The scientific consequence is that expectation errors on inflation are important in determining the level of unemployment. Friedman reiterates this point and then draws the policy implications:

> There is always a temporary trade-off between inflation and unemployment; there is no permanent trade-off. The temporary trade-off comes not from anticipated inflation per se, but from unanticipated inflation, which generally means from a rising rate of inflation. . . . A rising rate of inflation may reduce unemployment, a high rate will not.[42]

As a result, there could be no stable or predictable Phillips curve trade-off. The policy implications and social implications were equally clear. If policy makers, for example, attempted to reduce the existing rate of unemployment, the result would be more volatile swings in monetary policy and, by implication, in output, prices, and unemployment. Indeed, such policies would eventually be self-defeating and would instead create a combination of higher unemployment and higher inflation, or what came to be known as stagflation.

In the last 20 years, this new theory has been the dominant view held by policymakers and policy advisors of both political parties in the United States. While some can ascribe the elimination of stagflation in the United States to good luck, there is also evidence that scientifically informed policy practices were a factor.[43] Other western industrialized nations have followed suit.

All of this is not intended to applaud the discipline of economics at the expense of political science. It is simply to note that scientific progress in economics is testimony to the power of merging formal and empirical analysis. Economists have made numerous contributions in developing theories robust enough to guide policy actions. Still, many issues in the macroeconomic

policy area remain unresolved. Perhaps the most interesting to political scientists are those that center on the interaction of the public, elected officials, policy making institutions, and the ways in which these interactions can be structured to yield desirable policy responses. Political science can make progress in solving the conundrums that these interactions pose by constructing and progressively refining models that have well-identified relations with a direct empirical link.

Testing strategic interaction in international relations

A third example linking formal modeling with empirical analysis explores various aspects of relevance to the conduct of foreign policy and the use of military force. In particular, Curtis Signorino examined "the machinations of state leaders trying to achieve their foreign policy goals through sometimes peaceful but often violent means."[44]

Of central importance are the strategic behaviors and the strategic interactions between the competing states. As Signorino points out, "States do not act in a vacuum. Decisions to engage in arms production, enter into alliances, and go to war are not independent of the expected behavior of the other states in the system. Their calculations are based on what they expect other nations are currently planning to do or how they may respond to particular actions."[45]

For this type of research question, where strategic interaction is involved, game theoretic modeling offers important attributes. Additionally, Signorino was interested in linking game theoretic predictions with empirical tests. Signorino adopted an EITM approach with the following justification:

> Because of this emphasis on causal explanation and strategic interaction, we would expect that the statistical methods used to analyze international relations theories also account for the structure of strategic interdependence. Such is not the case. This article is an attempt to remedy that—to build a bridge between international conflict models and our statistical testing of these models.[46]

To link the formal game theoretic model with an empirical test, Signorino used Quantal Response Equilibrium (QRE) analysis. Originally developed by Richard McKelvey and Thomas Palfrey,[47] QRE allows for the agents in the game to make the best possible response (in terms of utility) to each other, and to make imperfectly informed decisions where the errors in the decisions have a distribution.

This second feature is of some importance. The inclusion of errors, while not only consistent with the idea of agents having bounded rationality (which is assumed), also allows for feasible statistical estimation.[48] Signorino reinforces the importance of this attribute in the following way:

> Specification of the distribution of those "errors" provides a statistical model of equilibrium, since it allows for nondegenerate (i.e., nonbinary) choice probabilities to be derived for the strategies players will choose. Hence, it allows for the derivation of nondegenerate probabilities for the outcomes of the game. Using a statistical equilibrium concept such as the QRE, one can derive the statistical version of a model of conflict in extensive form that directly incorporates the structure of the strategic interaction.[49]

Signorino did a series of tests to determine if QRE improved predictive accuracy. One test summarized here centers on his comparison of a strict empirical model (Logit) with data (generated via Monte Carlo) based on QRE. With his Monte Carlo experiments, he found that when the empirical model fails to incorporate strategic behavior, the statistically significant results give incorrect policy advice. Signorino pointed out the policy ramifications of following a naive, strictly empirical approach:

> Substantively, the estimates imply that an increase in military capabilities decreases the probability of war, an increase in assets increases the probability of war, and two democracies are unlikely to go to war. No doubt the researcher would note the counterintuitive finding that another nation's increase in military power actually decreases one's own probability of going to war with that nation. Since it is unrealistic to suggest that nations divest themselves of their assets, the analyst's policy prescription would be that nations should increase their military capabilities as much as possible: More military power leads to less war for everyone.[50]

However, if this policy prescription is followed it would increase the probability of war (and not reduce it) because the raw empiricism does not capture the true underlying probability of war in the Monte Carlo analysis.

Signorino also extended his Monte Carlo "experiments" to other more sophisticated models that include "balance of power" concerns or whether two nations jointly value war. He found that these improvements in sophistication, although estimated without consideration of strategic interaction, also provide poor policy advice. He concluded this portion of the analysis with the following thoughts that pertain to the issue of blindly using strictly empirical procedures and ignoring the richness a formal model can provide. This choice contributes to noncumulation:

> [M]ore troubling are the highly significant results in each case, which would be interpreted by the typical researcher as supporting one model or another. Hence out of a single data set, support could be "found" for a number of different theories of international conflict—all of which are wrong.[51]

Regulatory policy delay

Daniel Carpenter's research on regulatory policy offers a different type of policy relevant example where EITM is undertaken. Carpenter poses the question, "Why do bureaucrats delay?" Why do regulatory choices made under identical administrative procedures exhibit highly varying decision times? He studied the histories of 450 new drugs ("new chemical entities") reviewed by the United States Food and Drug Administration (FDA) between 1977 and 2000. Guided by an optimal stopping model—how can we predict the time horizon for approval of drugs?[52]—Carpenter gathered detailed information about the incidence and severity of the disease that the medication was intended to treat and tallied the number of existing approved drugs for that condition. He supplemented these data with information about the support and lobbying groups that worked on behalf of disease sufferers. Carpenter also assembled data on other factors that might be associated with drug approval, from the political makeup of Congress to how often the disease was mentioned in the media.

Carpenter found differences in the review time. Moreover, he discovered that factors that should not, given the FDA's mandate, make a difference in review time seemed to matter a great deal. For example, when Carpenter counted how many *Washington Post* stories mentioned a specific disease in any given year, and then computed how quickly new drugs for that disease won FDA approval, he found that the frequency of disease mentions seemed to relate directly to approval time for specific drugs. This occurred regardless of the seriousness of the ailment, its incidence in the population, the cost of treating it, the availability of other medicines targeting the disease, and other variables that measured the potential value of the drug.

Along with the influence of the print media, Carpenter also reveals equity issues that should concern policy makers. His model and his empirical tests together produce a set of interesting research findings with important social implications. The interaction between the optimal stopping model and his empirical results illuminate not just the degree of bureaucratic delay, but also the forces responsible for it. Consequently, this research not only contributes to an understanding of the linkages between the causes of, and implications of, delay in the drug approval process, but also provides findings for policymakers entrusted with formulating health policy.[53]

The Promise of Established Technical-Analytical Competence

If the goal is to build and test models with which we can establish causal relations, then current research and teaching practices will have to change. To accomplish this, a standard of technical and analytical competence must be established to augment extant research and teaching procedures. Only after a general standard of competence has been developed will the needed improvements be self-enforcing and commonplace. In this new research environment, formal models will respect data, proverbs will not be mistaken for theory, and one will no longer encounter articles filled with data mining and garbage-can empirics filtered by jerry-built statistical patches. Probably the most important indicator in knowing we have arrived will be when the letters E-I-T-M are no longer needed as part of the training for political scientists. At that point, it will be second nature to combine theoretical and empirical analysis for teaching and research.

How will these changes come about? As noted above, changes are already occurring.[54] Researchers in political science are beginning to develop designs that hold the promise of linking formal theory and empirical analysis in policy and socially relevant ways.[55]

Better technical and analytical training can be helpful. Innovation can be further accelerated through collaborations across subfields and disciplinary boundaries, with an added emphasis on ending separation between qualitative and quantitative research. In the various subfields of political science there is potential for collaboration between those who do case studies and/or study history and culture and those who wish to combine formal and empirical work.

Returning to the tasks that policy makers face, we think one of our functions as political and social scientists is to provide science-based advice that merits the utmost confidence in situations where the consequences of mistakes can be serious. For policymakers to have confidence in this advice, they need some assurance that a body of knowledge has been accumulated that identifies causal relations by rigorous means. Will this transformation in political science be difficult? Yes. Will it be frustrating and time consuming? Certainly. However, the scientific and social benefits gained from the EITM-induced cumulation of political science knowledge will be well worth the effort.

Appendix

A Markov Chain Model of Effects of Intergenerational Transfer of Voter Behavior

Markov chain analysis shows that even if two countries differ greatly in the extent to which voters identify with a party, if the rates of transferring identifications intergenerationally are the same, then over time the two countries will converge to the same party identification level.[56]

Let us assume the 80 percent and 50 percent rule:

- 80 percent of those whose fathers identified with a party develop party identification.
- 50 percent of those whose fathers do not have party identification develop party identification.

Assume further that party identifiers have the same number of children as non-identifiers. If 30 percent of the population of country A identifies with a party, and 90 percent of the population of country B identifies with a party, in the next generation we would see by application of a Markov chain:

$$(0.8 \times 30 \text{ percent}) + (0.5 \times 70 \text{ percent}) = 59 \text{ percent of country A having identification,}$$

and

$$(0.8 \times 90 \text{ percent}) + (0.5 \times 10 \text{ percent}) = 77 \text{ percent of country B having identification.}$$

Then in the next generation, we would see:

$$(0.8 \times 59 \text{ percent}) + (0.5 \times 41 \text{ percent}) = 67.7 \text{ percent of country A having identification,}$$

and

$$(0.8 \times 77 \text{ percent}) + (0.5 \times 23 \text{ percent}) = 73.1 \text{ percent of country B having identification.}$$

Note the predicted difference has shrunk from 60 percent to 5.4 percent with the passage of the generations.

Notes

1 Riker 1962, viii.

2 The views presented in this paper are those of the authors and do not necessarily reflect official National Science Foundation policy.

3 Report of the APSA Ad Hoc Committee 2000, 1.

4 Powell 1999.
5 Wagner 2004.
6 Achen states that these empirical models (or practices) "are too often long lists of independent variables from social psychology, sociology, or just casual empiricism, tossed helter-skelter into canned linear regression packages." Achen 2002, 424. He credits Anne Sartori for the term "garbage-can regressions."
7 Formal analysis—or formal modeling—includes, among other things, deductive modeling in a theorem and proof presentation or computational modeling that requires the assistance of simulation. Empirical analysis usually involves either data analysis using statistical tools or case studies.
8 This argument was the basis for the NSF-sponsored workshop on Empirical Implications of Theoretical Models, July 9–10, 2001. The EITM Report is available at www.nsf.gov/sbe/ses/polisci/eitm_report/start.htm. Copies of the report are also available upon request from the NSF Political Science Program. Address requests to: EITM Report, Political Science Program, National Science Foundation, Suite 980, 4201 Wilson Boulevard, Arlington, VA 22230.
9 Wagner 2001, 3.
10 See Coase 1960; Duverger 1954.
11 Allais 1988, 245.
12 Lin 2004.
13 Brady and Collier (forthcoming).
14 Koremenos, Lipson, and Snidal (2001a; 2001b). We view the debate between Elster (2000) and Bates et al. (2000) as a positive sign and consistent with the direction of Koremenos et al.
15 Brady and Collier (forthcoming).
16 The NSF Political Science Program provides support to numerous infrastructure-building activities related to these approaches. In addition to standard research grants, the Program provides funds for training and workshops. See these sites for further information: Qualitative (http://www.asu.edu/clas/polisci/cqrm/index.html); Quantitative (http://polmeth.wustl.edu/conferences.html); and EITM (http://www.cbrss.harvard.edu/eitm.htm; http://wc.wustl.edu/eitm/; http://www.isr.umich.edu/cps/eitm/eitm.html; http://www.poli.duke.edu/eitm/).
17 Akerlof 2002, 412.
18 Key 1954, 184.
19 Simon 1946, 53.
20 For a recent treatment of many of these issues see Wagner 2004.
21 Riker 1982, 4.
22 Ibid., 5.
23 Key 1954, v–vi.
24 See, for example, Johnston and DiNardo 1997.
25 Achen 2000, 143.
26 Lin 1999.
27 Brady (forthcoming); Lott 2000; Lott 2001.
28 Brady (forthcoming), 283.
29 Brady (forthcoming), 284.
30 Achen 2002, 445.
31 Friedman 1991, 48.
32 Ibid., 49.
33 This example is adapted from Shively 1990.
34 Converse and Dupeux 1962.
35 Shively 1990, 17.
36 Converse 1969.
37 Shively 1990.
38 Ibid., 20.
39 Gerber and Green 1998.
40 Phillips 1958.
41 Friedman 1968 and Phelps 1968.
42 Friedman 1968, 11.
43 Ahmed et al. 2002.
44 Signorino 1999, 279.
45 Ibid.
46 Ibid.
47 McKelvey and Palfrey 1995; McKelvey and Palfrey 1996; McKelvey and Palfrey 1998.
48 In more technical language, QRE avoids the use of nonzero probabilities, which makes, for example, likelihood statistical procedures impossible. For details, see Signorino 1999.
49 Signorino 1999, 282.
50 Ibid., 286.
51 Ibid., 288.
52 Dixit 1993.
53 Carpenter 2002.
54 See Morton 1999.
55 See Mebane and Sekhon 2002; Sartori 2003; Schultz 2001.
56 Shively 1990.

References

Achen, Christopher. 2000. Warren Miller and the future of political data analysis. *Political Analysis* 8:2, 142–6.

———. 2002. An agenda for the new political methodology: Microfoundations and ART. *Annual Review of Political Science* 5, 423–50.

Akerlof, George A. 2002. Behavioral macroeconomics and macroeconomic behavior. *American Economic Review* 92:3, 411–33.

Ahmed, Shaghil, Andrew Levin, and Beth Anne Wilson. 2002. Recent U.S. macroeconomic stability: Good policies, good practices, or good luck? International Finance Discussion Paper, Board of Governors of the Federal Reserve System, 730, July.

Allais, Maurice. 1988. An outline of my main contributions to economic science. Nobel Lecture, 9 December. Ecole Nationale Supérieure des Paris et Centre National de la Recherche Scientifque-France.

Bates, Robert, Avner Greif, Margaret Levi, Jean-Laurent Rosenthal, and Barry R. Weingast. 2000. The analytic narrative project. *American Political Science Review* 94:3, 696–702.

Brady, Henry. Forthcoming. Data-set observations versus causal-process observations: The 2000 U.S. presidential election. In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, ed. Henry Brady and David Collier. Boulder, Colo., and Berkeley, Calif.: Rowman and Littlefield, and Berkeley Public Policy Press, 279–84.

Brady, Henry, and David Collier, eds. Forthcoming. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Boulder, Colo., and Berkeley, Calif.: Rowman and Littlefield, and Berkeley Public Policy Press.

Carpenter, Daniel P. 2002. Groups, the media, agency waiting costs, and FDA drug approval. *American Journal of Political Science* 46:3, 490–505.

Coase, Ronald. 1960. The problem of social cost. *Journal of Law and Economics* 3:1, 1–44.

Converse, Philip E. 1969. Of time and partisan stability. *Comparative Political Studies* 2 (July), 139–71.

Converse, Philip E., and Georges Dupeux. 1962. Politicization of the electorate in France and the United States. *Public Opinion Quarterly* 26:1, 1–23.

Dixit, Avinash. 1993. *The Art of Smooth Pasting*. Chur, Switzerland: Harwood Academic.

Duverger, Maurice. 1954. *Political Parties: Their Organization and Activity in the Modern State*. New York: Wiley.

Elster, Jon. 2000. Rational choice history: A case of excessive ambition. *American Political Science Review* 94:3, 685–95.

Empirical Implications of Theoretical Models Report. 2002. Political Science Program, National Science Foundation, Directorate of Social, Behavioral, and Economic Sciences. Arlington, Va.

Friedman, Milton. 1968. The role of monetary policy. *American Economic Review* 58:1, 1–17.

———. 1991. Appendix: A cautionary tale about multiple regressions. *American Economic Review* 81:1, 48–9.

Gerber, Alan, and Donald P. Green. 1998. Rational learning and partisan attitudes. *American Journal of Political Science* 42:3, 794–818.

Johnston, Jack, and John DiNardo. 1997. *Econometric Methods*. New York: McGraw-Hill.

Key, V. O., Jr. 1954. *A Primer of Statistics for Social Scientists*. New York: Thomas Y. Crowell.

Koremenos, Barbara, Charles Lipson, and Duncan Snidal. 2001a. Rational design: Looking back to move forward. *International Organization* 55:4, 1051–82.

———. 2001b. The rational design of international institutions. *International Organization* 55:4, 761–99.

Lin, Tse-min. 1999. The historical significance of economic voting, 1872–1996. *Social Science History* 23:4, 561–91.

———. 2004. Wittgenstein and mathematical political theory. Typescript, University of Texas, Austin.

Lott, John R., Jr. 2000. Gore might lose a second round: Media suppressed the Bush vote. *Philadelphia Inquirer*, 14 November.

———. 2001. Documenting unusual declines in Republican voting rates in Florida's western panhandle counties in 2000. Typescript, Yale University Law School.

McKelvey, Richard D., and Thomas R. Palfrey. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10:1, 6–38.

———. 1996. A statistical theory of equilibrium in games. *Japanese Economic Review* 47:2, 186–209.

———. 1998. Quantal response equilibria for extensive form games. *Experimental Economics* 1:1, 9–41.

Mebane, Walter, and Jasjeet Sekhon. 2002. Coordination and policy moderation at midterm. *American Political Science Review* 96:1, 141–57.

Morton, Rebecca B. 1999. *Methods and Models: A Guide to the Empirical Analysis of Formal Models in Political Science*. Cambridge: Cambridge University Press.

Phillips, A. W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 25:100, 283–99.

Phelps, Edmund. 1968. Money wage dynamics and labor market equilibrium. *Journal of Political Economy* 76:4, part 2, 687–711.

Powell, Robert. 1999. *In the Shadow of Power: States and Strategies in International Politics*. Princeton: Princeton University Press.

Report of the APSA Ad Hoc Committee on the National Science Foundation. 2000. American Political Science Association, 1527 New Hampshire Avenue, NW Washington, D.C. 20036.

Riker, William H. 1962. *The Theory of Political Coalitions*. New Haven: Yale University Press.

———. 1982. *Liberalism Against Populism: A Confrontation Between the Theory of Democracy and the Theory of Social Choice*. San Francisco: W. H. Freeman.

Sartori, Anne E. 2003. An estimator for some binary-outcome selection models without exclusion restrictions. *Political Analysis* 11:2, 111–38.

Schultz, Kenneth. 2001. *Democracy and Coercive Diplomacy*. Cambridge: Cambridge University Press.

Shively, W. Phillips. 1990. *The Craft of Political Research,* 3d ed. Upper Saddle River, N.J.: Prentice Hall.

Signorino, Curtis. 1999. Strategic interaction and the statistical analysis of international conflict. *American Political Science Review* 93:2, 279–97.

Simon, Herbert. 1946. The proverbs of administration. *Public Administration Review* 6:1, 53–67.

Wagner, R. Harrison. 2001. Who's afraid of "rational choice theory"? Typescript, Department of Government, University of Texas, Austin.

———. 2004. *War and the State: An Introduction to the Study of International Politics*. Unpublished manuscript, University of Texas, Austin.