Modeling Practices to Ensure a Science of Science Policy*

Jim Granato
Center for Public Policy
University of Houston

Frank Scioli
National Science Foundation

Michelle Costanzo
University of Maryland

"The relationship between the growth of science, technological development, and social change has always been a core concern in studies of the long-term transformation of Western societies … In the past two decades especially, some of the focus on the relations between science, technology, and society has turned in more policy-oriented directions. Controversy has arisen over the quality and the dissemination of knowledge … and over the nature of public control and regulation intended to provide benefits of scientific and technological progress while minimizing the hazards." (Gerstein et al. 1988: 144).

I. The Challenge

The preceding quote sketches an ongoing challenge for establishing a science of science policy. We describe *the science of science policy* as the use of scientific methods for assessing investments in research and development.

For obvious reasons the challenge to improve the science of science policy occupies an important place in academia and government. In fact, a recent statement from the Office of Science and Technology Policy, as part of the American Competitiveness Initiative (ACI), includes this challenge as fundamental to the long-term objectives of the ACI. The ACI document asserts that it is essential to have:

"Federal investment in cutting-edge basic research whose quality is bolstered by merit review and that focuses on fundamental discoveries to produce valuable and marketable technologies, processes, and techniques." (Domestic Policy Council 2006: 1).

The science of science policy is a complex endeavor primarily because the policy analyst is rarely able to analyze systematically a complex policy question; especially if there are several options available for achieving a

---

* This paper is a revised version of the paper prepared for presentation at the Atlanta Conference on Science and Technology Policy 2006, US-EU Policies for Research and Innovation. Technology Square, Georgia Institute of Technology, Atlanta, Georgia, May 18-20. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Brian Humes and Mark Jones for their comments and criticisms.

particular set of objectives.[1] What can add to the complexity is that in many cases the science of science policy involves the incorporation of social, behavioral, and economic processes into a policy model. Under these circumstances, the tools/metrics of social science are necessary if the policy analyst (as evaluator) wants to be able to assess the effects of a specific policy on achieving its intended objectives.[2]

To be sure, policy maker discretion, for example, --- in economic policy --- has been influenced by evolving social scientific research practices (and their attendant metrics), particularly when these practices are cumulative. In these instances, policy maker goals or targets can receive greater or lesser weight based not only on current conditions (that policy makers tend to be most familiar with) but also on what policy analysts may predict is the optimal course of action. Yet, it is not always clear that cumulative practices are being followed. When noncumulative practices are used policy makers are better off relying on using their experience and "common sense" because the policy analyst's research will be so flawed as to only confuse things.

Against this background, this paper speaks to both academic and government communities. For academics, our objective is to encourage scientific examination (via cumulative scientific practices) of the most accurate representation of phenomena of interest and how policy can influence such phenomena. It is critical, in this regard, that students of public policy utilize their academic training and knowledge in order to possess the requisite methodological tools for careful analysis of substantive science policy. On the other hand, we also recognize that policy makers --- which include policy advisors to the executive administration --- are focused on the development of public policy that is effective in meeting the needs of society. They are entrusted with taking the right policy action at the right time. But, it is also in a policy makers interest to make use of valid science policy in order to augment their experience and common sense.

In short, both communities have much to gain from the effort to improve upon existing social science research practices and metrics. Along these lines, John Marburger commented in the 2002 AAAS Colloquium on Science and Technology Policy that:

"Scientists do, of course, make judgments all the time about promising lines of research . . . it makes sense for the world's largest sponsor of research, the U.S. government, to want to make such choices as wisely as the most productive scientists do…But is it possible to decide rationally when to enhance or to terminate a project if we do not possess a way of measuring its success?"

While the science of science policy is a very broad topic, we will narrow our focus to a framework for model building and model testing that can assist in furthering the development of a cumulative science of science policy. We first discuss how model building and

[1] See Feldman et al. (2003).

[2] Metrics can be defined as measures as well as statistical and mathematical analysis.

testing practices --- the development of metrics --- can benefit from closer collaboration between academics and policymakers. Next, since modeling and testing practices can involve many things, we narrow the discussion to applied statistical practices as they relate to criteria for scientific cumulation.[3] From the discussion of various applied statistical practices, and taking into account the needs and expertise of these two communities, we provide a framework for the establishment of procedures where the science of science policy (across disciplines and policy arenas) can be strengthened. In this section we also outline current applied statistical practices that discourage scientific cumulation. The fifth section describes how this framework encourages a shift in statistical indicator(s) emphasis. The next section then explores the societal consequences of our framework. The paper concludes with a discussion of short-term and long-term implementation strategies of the framework.

[3] While (applied) statistical/econometric analysis is the primary modeling practice we focus on in this paper, we will include formal analysis (formal modeling). Formal analysis refers to deductive modeling in a theorem and proof presentation or computational modeling that requires the assistance of simulation. Applied statistical analysis involves data analysis using statistical tools. Throughout this paper we will use the words analysis and modeling interchangeably.

II. The Case for Blurring the Distinction Between Academics and Policymakers

Although the academic and policymaking communities have different ways of looking at the world, we contend that it would be a mistake to treat these two groups separately. Academics in the social sciences establish research strategies for differentiating between the effectiveness of various policy alternatives. On the other hand, policy makers do the same thing except that their "data" are their experiences and day-to-day expertise.

Because of this unity of purpose, both groups are fundamentally interested in developing research models that allow a clear identification of policy objectives and then providing a set of metrics for assessing whether the policy objectives are, in fact, linked to specific policy outcomes. This is classic program evaluation methodology (relating policy formulation and policy implementation to policy outcomes) and is an entire academic field in and of itself (Patton (2002)), Rossi, Lipsey, and Freeman (2004)).

In some policy areas this is less difficult to do than others. For example, it may be possible to develop research metrics for assessing whether clean air policy results in a reduction of particulate matter in the atmosphere. Likewise, it is possible to assess whether fatalities from certain diseases can be lowered in patients undergoing treatments in specific clinical trials. Although we acknowledge that these judgments are not easy to establish unless rigorous experimental conditions are imposed nonetheless, the development of specific metrics linking policies and outcomes is still feasible.

For policy areas, not subject to experimental design and experimental control, it is far less easy to develop the metrics.[4] This is due to the difficulty of establishing quantifiable outcomes. Academics address this by attempting analysis of both formulation and implementation of policy and then attempting to establish measures which quantify the "outcomes" of specific policies. Policymakers may, or may not, use this analysis when choosing among policy alternatives. For example, calculating return on dollar investment, independent of social effects, is one common research strategy. However, because we argue for an integration of policy formulation, implementation, and evaluation strategies (within a social scientific framework), we think policy makers are more likely to make systematic policy choices if they are provided with evaluations that are broader than simply looking at data based on return on dollar investment. This is particularly important for the budget formulator when choosing among policy alternatives so that budget dollars are spent on effective policies.

Along with encouraging alternative criteria for evaluation, the integration of academic and policymaker expertise is especially important for avoiding policy errors. Economic policy in the United States provides a useful case study in this regard. During the 1930s, policymakers failed to engage in stimulative monetary policy when extant economic theory, the province of policy analysts, dictated that course of action (Friedman and Schwartz (1963)). The Great Depression was sustained as a result. At the other extreme, the stagflation of the 1970s occurred because academic policy analysts predicted that policymakers could "control" unemployment and inflation simultaneously through stabilization policy. However, real-time experiences indicated that repeated attempts to lower unemployment lead to both higher inflation and higher unemployment (Granato and Wong (2006)).[5]

With the establishment, and integration, of reliable and valid social science metrics, policy makers will have the confidence to accept recommendations from policy analysts. Nevertheless, we

[4] In work supported by the National Science Foundation to construct datasets to study the policy-making process, the Principal Investigators note that extant datasets suffer from major reliability problems: Frank Baumgartner, Bryan Jones, and John Wilkerson, Collaborative Research: Database Development for the Study of Public Policy, National Science Foundation #0111443, Policy Agendas Project: http://www.policyagendas.org.

[5] Examples of the interplay between social scientific research and policymaker experience are found in Meyer (2004). For example, on the relation between productivity, wages, and inflation he states:

"Economic theory tells us that a leap in productivity will raise wages in the long run. But experience tells us that wages are not initially much affected. As a result, in the short term, an increase in productivity tends to lower the cost per unit of output. This, in turn, will generally push prices down." (p. 126).

hasten to add that the fusion of scientific policy research with practical day-to-day policy activity cannot totally eliminate harmful policy choices, given the uncertainties of certain scientific relations and the pressure of making real time decisions. The point is that there has to be a dialogue between policy analysts and policy makers, while maintaining an appreciation of the difference between what the models predict and reality.

Alan Greenspan (2004) suggests that peaceful coexistence is possible in the conduct of policy:

"In designing strategies to meet our policy objectives, we have drawn on the work of analysts...A critical result has been the identification of a relatively small set of key relationships that, taken together, provide a useful approximation of the economy's dynamics... However,...our knowledge of the important linkages is far from complete and, in all likelihood, will always remain so (p. 5)...For such judgment, policymakers have needed to reach beyond models to broader --- though less mathematically precise --- hypotheses about how the world works." (p. 7).

III. Modeling Practices that Lead to Scientific Cumulation

"We need econometric models that encompass enough variables in a sufficient number of countries to produce reasonable simulations of the effect of specific policy choices." (Marburger 2005: 1087).

The science of science policy will need some fundamentally new modeling practices that will not only build upon existing metrics but, when possible, create new ones. But, what scientific criteria should guide this endeavor? Replacing old statistical techniques with new statistical techniques that supposedly have more statistical "power" is not enough to uncover causal mechanisms.

Nor is it enough to use predictive accuracy as the primary basis for building a cumulative science. Predictive accuracy is very useful in that it provides model reassessment and also can promote a dialogue between policy makers and policy analysts. However, in order to ascribe responsibility for predictive (in)accuracies we need more than guesses about what is in a "black box." Specifically, metrics for predictive accuracy need to be part of an overall modeling strategy that makes a transparent link between models and tests through valid hypothesis testing. This transparency and testing is essential to the process of model reassessment.

To guide in the development of transparent linkages between models and tests, we look to the basics of scientific inquiry. One distinguishing feature of scientific practice is that hypotheses must be refutable. Refutation, or *falsifiability*, is essential for building a cumulative science since hypothesis testing can provide feedback on the internal consistency of a model (illuminating internal workings of a model) before predictive assessments can proceed. Further, there are already useful metrics and standards in place for hypothesis support or refutation (i.e., Type I and Type II errors).

While valid predictive assessments are best assured when linked with valid hypothesis tests, there also must be some basis for determining hypothesis test validity. Among other things valid hypothesis tests rely on the ability of the policy analyst to derive cause and effect mechanism(s) that are *identified* and *invariant*. The terms *identification* or *identifiability*, when applied to the science of science policy, means the development of a model that possesses sufficient information so that a unique and valid inference can be drawn from a parameter of interest. The terms *invariant* or *invariance*, when applied to econometric models, centers on whether a relation (signified by a parameter) remains constant in the face of a treatment (or policy) shift.

Since it cannot be assumed that the science of science policy can always be studied in a rigorously controlled environment, models can only be supported or falsified probabilistically (i.e., chance). This is not the ideal way to establish effects. As was noted earlier, the most effective strategy would be to conduct true experiments, but it may not be feasible to do so. Thus, what is needed is a quantitative indicator, which over time will become an important point of valid hypothesis testing (avoiding Type I and Type II errors), and also will provide information on identification and invariance.

In this regard, few indicators exceed the application and importance of a t-statistic. A t-statistic is defined as the ratio, (b/(s.e.(b))). The t-statistic allows probabilistic statements (i.e., statistical significance) and its numerator (b) contains the information pertaining to identification and invariance. In sum, this identity contains the attributes for exploring valid policy simulations.[6]

The use and attributes of the t-statistic are enhanced when both formal and applied statistical modeling techniques are linked to each other. In other words, identified and invariant relations (as seen in the (b)) can be derived and tested for when there is a transparent linkage between formal and empirical analysis.

There is good reason for this linkage since formal models force clarity about assumptions and concepts; they ensure logical consistency, and they describe the underlying mechanisms that lead to outcomes. Likewise, applied statistical models provide generalizations and rule out alternative explanations through multivariate analysis. Applied statistics assist in distinguishing between causes and effects, allow for reciprocal causation, and also assist in understanding the relative size of the effects.

To recap briefly, we are arguing that the components of the metrics for a new science of science can include (but are not limited to) the following:

1) Predictive accuracy that is closely tied to rigorous hypothesis testing

[6] In addition to the concept of statistical significance is the concept of s*ubstantive significance* (the size of the numerator (b)). Substantive significance is of great importance since it demonstrates how borderline (highly) significant t-statistics can be associated with big (small) effects of the independent variable on the dependent variable.

and therefore allows for valid model reassessment.

2) Hypothesis testing that makes use of a t-statistic and is based on identified and invariant parameters.

3) Linking formal models to applied statistics so that their mutually reinforcing features can illuminate identified and invariant parameters and provide a transparent linkage between model and test.

IV. A Proposed Framework

With these components in place, the basic idea of this framework is to take what social scientists know about theoretical and applied statistical concepts, provide a rigorous basis for these concepts through the use of their respective analogues, and then merge these theoretical analogues with the applied statistical analogues.[7]

What would emerge is a road map for others to modify, correct, or follow. More importantly, one could provide a transparent linkage between a theory and test. This is not to say the model is correct. Instead, it involves meeting a minimal requirement that the theory and test are related and, therefore, refutable.

Our approach/framework in some respects is not new since it relies (in part) on the work of the Cowles Commission --- as presented in the 1930s and 1940s.[8] The Cowles Commission was a group known for establishing conditions in which structural parameters could be determined, and causal mechanisms could then be identified by using these structural parameters.[9] But, unlike the

---

[7] A *concept* can be thought of as an abstract or general idea inferred or derived from specific instances. An *analogue* can be thought of as a device in which a concept is represented by continuously variable --- and measurable --- quantities.

Behavioral concepts include but are not limited to *expectations, learning, and social interaction*. Behavioral analogues include but are not limited to the use of *conditional expectations*, and *adaptive learning*, and the use of *system level analysis*. Statistical concepts include but are not limited to *persistence*, *measurement error*, or *simultaneity*. Statistical analogues include but are not limited to *autoregressive processes, error-in-variables regression*, and *simultaneous equation estimation*.

[8] For further background on the Cowles Commission consult: http://cowles.econ.yale.edu/.

[9] We adopt Heckman's (2000: 59) terminology below:

"Structural causal effects are defined as the direct effects of the variables in the behavioral equations...When these equations are linear, the coefficients on the causal variables are called structural parameters (emphasis added), and they fully characterize the structural effects."

Heckman also notes there is some disagreement about what constitutes a structural parameter. The disagreement centers on whether one uses a linear model, a non-linear model or, more, recently a fully parameterized model. In

Cowles Commission, our framework also links formal and applied statistical analogues to assist in this identification process. We will discuss the Cowles Commission at greater length later in this paper.

While scientific cumulation is based on many things, we maintain our theme of requiring that the science of science policy (and policy simulation) be based on the search for identifying invariant causal mechanisms. Since we accept the Cowles Commission's focus on structural parameters as an important way to demonstrate specific cause and effect, we place particular emphasis (when using the t-statistic) on the numerator (b) as opposed to practices that emphasize the denominator (s.e.(b)).

We can summarize these issues (and some others) in Figure 1. For purposes of contrasting what can be done to improve upon In this figure we incorporate what we believe are nonfalsifiable practices that provide no useful metrics and can harm science policy development.

(Figure 1 About Here)

A1. Scientific Accumulation: On the Y-axis is the criterion of scientific cumulation. We assert that one way to determine scientific cumulation is a significant t-statistic, which (to repeat) is defined as the ratio, (b/(s.e.(b))). 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 is not usually the case. Instead, the focus of current policy research is usually on the size of the t-statistic and whether one can get significant results.

A2. Identification/Invariance: On the X-axis is the "identification/invariance" criterion.[10] Recall we have asserted that cumulative science cannot progress when research practices are non-falsifiable and where some assurance is given that the parameter(s) --- the b's --- reflect the effect of the independent variable(s) in question. This provides guidance as to whether either type of error noted above is avoided.[11]

A3. Predictive Accuracy: A final issue in the process of linking model to test is how accurate the model predicts. There are numerous measures for predictive accuracy, but no less important is that the use of these measures fosters model reassessment, promotes a dialogue between policy makers and policy analysts on ways to improve a model, and ultimately contributes to scientific cumulation.

---

the latter case, structural parameters can also be called "deep" to distinguish between "*the derivatives of a behavioral relationship used to define causal effects and the parameters that generate the behavioral relationship*." (p. 60).

[10] The two criteria can be listed on separate dimensions, but for simplicity we combine them since they both are necessary for valid policy analysis.

[11] Kuhn (1969) has noted that scientific revolutions are noncumulative developmental episodes. We agree, but also add that developmental episodes also need to have falsifiable properties.

Despite these useful scientific properties, it is important to identify current methodological practices that use the t-statistic, but do not make proper use of the information in the (b). Through the manipulation of standard errors, (s.e.(b)), instead of isolating and identifying structural parameters, the t-statistic loses its scientific merit. These practices contribute to noncumulation and spurious conclusions.

The risk for policy errors is heightened when metrics and criteria, unrelated to identifying invariant mechanisms, dominate. Recognition of the non-scientific threat posed by this methodological practice originates with the econometric approach (and criteria) used by the Cowles Commission, described earlier. It was the Cowles Commission that explored the differences between structural and reduced-form parameters. Conditions for identifiability were introduced to aid in this differentiation and today this method is part of standard texts in econometrics.[12]

The Cowles Commission's contributions rested, in part, on a scientific vision that involved merging formal and applied statistical analysis. The basis for this linkage was the idea that random samples were governed by some latent and probabilistic law of motion (Haavelmo (1944), Morgan (1990)). Then using this "law", formal models, when related to an applied statistical model, may be interpreted as creating a sample. Within this framework, a well-grounded test of a theory could be accomplished by relating a formal model to an applied statistical model and testing the applied statistical model. This methodological approach was seen, then, as a valid representation and examination of underlying processes in existence.

Although the Cowles Commission contributed to the development of scientific practices, today there are some who still employ a so-called technical method that is only loosely connected to valid hypothesis testing --- and Cowles Commission practice. In these cases, there is a strong reliance on statistically significant results with only little attempt to identify the precise origin of the parameters in question. Absent this identification effort, it is not evident where the model is wrong. In a different sense, these current practices are getting ahead of themselves by failing to establish ways to falsify the models.

Three harmful but common applied statistical practices exist: data mining, garbage cans, and the use of statistical weighting and patching (i.e, the use of Omega Matrices). These three practices are incompatible with falsification. Consequently, scientific cumulation is an unattainable goal. Why? These applied statistical practices lack overall robustness. Worse, they can be used to obscure some fundamental specification error. Again, referring to Figure 1 (below the X-axis), we summarize these practices. These practices are situated closest to the origin of the figure since they fail to identify invariant parameters and assist in valid hypothesis testing.

[12] Along with their work on structural parameters, the Cowles Commission also gave formal and empirical specificity to issues such as exogeneity and policy invariance (Morgan (1990), Heckman (2000: 46)).

B1. Data Mining: One practice that can achieve statistically significant results involves putting data into a statistical package with minimal theory. Regressions (likelihoods) are then estimated until either statistically significant coefficients, or coefficients the researcher thinks are important, are attained. This step-wise search has little relation to identifying causal mechanisms.

B2. Garbage Cans: A second practice, related to data mining, involves researchers, including, in a haphazard fashion, a plethora of independent variables into a statistical package and then obtaining significant results. Researchers using garbage can models rarely pay attention to potential confounding factors that could corrupt statistical inferences. Efforts to identify an underlying causal mechanism are also few and far between.

B3. Omega Matrices: Data mining and garbage-can approaches virtually are guaranteed to break down statistically through both inaccurate predictions and the existence of nonrandom error. The question is what to do when these failures occur. There are elaborate ways of using (error) weighting techniques to correct model misspecifications or to use other statistical patches that influence s.e.(b). In almost any intermediate econometrics textbook one finds a section (chapter) 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, ultimately the standard error noted above, is altered and the t-statistic is manipulated. In principle, there is nothing wrong with knowing the Omega matrix for a particular statistical model. 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 practice hampers scientific progress because it uses a model's mistakes to obscure flaws.

In addition to these flawed applied statistical practices, it should also be pointed out that some formal practices threaten the science of science policy.

B4. Formal Models that Fail to Respect Facts: Formal models can fail to incorporate empirical findings that would assist in providing a more accurate depiction of the relations that are specified. 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.

Taken as a whole, these research practices, all too common today, are inadequate for the task of scientific cumulation --- and building a science of science policy. Recall that we accept the Cowles Commission's focus on parameters and identification. An emphasis on parameters --- structural parameters --- provides transparent interpretation and valid hypothesis testing (minimizing Type I and Type II errors).

However, structural parameters are only part of what is necessary for valid hypothesis testing. It is equally important to include the behavioral traits such as public tastes, attitudes, expectations, learning, and the like. Absent these added behavioral concepts and analogues, formal and applied statistical modeling can be misleading because the shifts of variables in the system failed to be linked with the other variables and parameters that are specified. Invariance can be compromised in this situation.

The framework we propose is summarized inside the upper right quadrant of Figure 1. We start with a structural description of the system by adding behavioral concepts, their formal and applied statistical analogues, and then how to link them. This framework provides information to assist in identifying causal linkages. By identifying *behavioral* concepts and analogues, their shifts can now be modeled and linked to a test with the parameters reflecting changes in causal influences.[13]

C1. Relating Behavioral and Applied Statistical Concepts: We begin by showing how well known social, behavioral, political, and economic concepts can be related to well known applied statistical concepts. An important way to enhance the linkage is through one kind of qualitative analysis, the case study method (as well as policy maker experience). Case studies provide detailed information about the steps by which events occur and allow researchers to identify mechanisms that can be incorporated into a formal model. In fact, as we noted elsewhere[14], qualitative analysis and quantitative analysis contribute to cumulative knowledge when thought of, and used as, mutually reinforcing methods. Thus, the approach we advocate is to encourage and accelerate shared standards and multi-method approaches.

C2. Behavioral and Applied Statistical Analogues: To link concepts with tests, we need analogues. The use of analogues is probably the most important advance over and above the Cowles Commission research program. Analogues serve as analytical devices for modeling and predicting behavior and, therefore, provide a richer and more identifiable interpretation of the formal and applied statistical model.

C3. Link the Formal and Applied Statistical Analogues: We then link these formal and applied statistical analogues for the purposes of identifying parameters of interest (b's). The attributes of this linkage means shifts in behavior and predictive accuracy are intertwined. Examples of this type of mutually reinforcing linkage include, but are not limited to, linking the analogue for expectations (conditional expectations) with the analogue for persistence (autoregressive process) or measurement error (error-in-variables regression).

V. How Cumulative Practices Influence the Science of Science Policy: A Focus on Parameters

When we make use of analogues and focus on the relation between our formal-theoretical parameter(s) and the applied statistical parameter(s) valid hypothesis testing, improved predictive

[14] See Granato and Scioli (2003).

accuracy (via model reassessment), and cumulation are possible. Explicit emphasis on the parameters allows for greater likelihood of knowing what is being falsified. This approach accounts for several methodological problems and is a more robust metric than the current practices of data mining, garbage cans, and Omega matrices.

Having said this, it is also clear that the models can still be over parameterized. While this framework can work out the mechanism between the policy-treatment (parameter(s)) and the outcome(s) of interest, we recognize that over parameterization ("free" parameters) can have unaccounted for consequences for causal linkages. The concern here is that the potential shifts in the free parameters may undermine invariance. Lucas (1980) poses the problem and suggests a possible accommodation when he states:

"If this [free] parameter changes in reaction to changes elsewhere in the system..., there is no way to predict the nature of these responses short of experimenting with the system as a whole" (p. 712).

To account for this possibility, our framework must be extended to include ways "to fix" as many parameters in the system as possible. By "fix" we mean metrics must be developed or adopted that provide greater specificity on the range (i.e., magnitudes) of free parameters. This effort would also include whether all free parameters must be subject to some restriction.

The framework we propose links formal models and empirical data and is flexible enough to accomplish this task. The issue is to make appropriate use of additional formalization --- informed by empirical and contextual evidence.

VI. The Payoff: Societal Consequences

We argue that the framework outlined above, when adhered to, can yield significant and positive societal benefits. If policy makers can accept the analyses presented to them by policy analysts (who use metrics with identifiable and invariant attributes), then they --- the policy makers --- will be able to make informed judgments about the best way(s) to assess policy effectiveness. They will also be able to make tough budgetary choices concerning what expenditures have the potential for the greatest payoffs.

By way of example, the effect of this interaction between empirical (and contextual) evidence and formal theorizing can be seen in the way that policy makers formulate decisions about how to stabilize the business cycle. Over the past 50 years the volatility of business cycles has been reduced and the duration of economic expansions has increased in the United States (Granato and Wong (2006)).[15] These salutary economic events occurred at approximately the same time that quantitative methodologies (inspired in part by the Cowles Commission) emphasized and were judged on their ability to produce identifiable and invariant predictions.

[15] Since 1854 the three longest peacetime (or otherwise) economic expansions in the United States occurred ***after*** World War II.

Is this relation between quantification and prosperity a correlation? A very strong case can be made that this important benefit to society is in part a function of the systematic use of quantifiable models. These models, while having seen their share of criticism, have assisted policymakers by providing useful knowledge and creating a systematic scientific justification for policymaker actions.

VII. Conclusion: Short-term and Long-term Solutions

If policymakers and policy analysts are to improve upon current practices, then the challenge is to make better use of both experimental and non-experimental data to make more accurate predictions. To do this science policy must, for example, be based on models that fit the past but also can be simulated to give valid (and reliable) estimates of the effects of various policies. To meet this challenge we have discussed how modeling practices must emphasize identifying causal mechanisms that are invariant to outside shocks.

We think immediate changes in model assessment (consistent with the issues of identification and invariance) can be instituted with little cost in terms of re-tooling or adding further complications to presentation criteria. The benefits would be two-fold. First, the added retooling would allow consumers of such information (i.e., policymakers) to focus directly on the validity of any model result. Second, the changes in criteria will create a more demanding standard for policy analysts to meet. Any models that fail to meet these criteria will be disregarded for the inherent vagueness and lack of assurance in the validity of the results.

The criteria for existing models can include the following questions[16]:

1) What assurances are given in the analysis that the parameters are invariant?

2) What alternative arguments (variables) were considered and what selection criteria were used to reach the final model formulation?

In the long-term, however, more ambitious changes will be needed. The transformation of current science policy modeling practices will not occur quickly. Fundamental changes are needed in how and what modeling skills are taught. We have outlined a framework that seeks to merge both formal and empirical models. It is this transparency between model and test that will ultimately transform the science of science policy.

[16] The criteria would require some policymakers to re-tool but our collective experience in teaching these tools demonstrates that this can be accomplished with little more than a one-day workshop.

References

Domestic Policy Council. 2006. *American Competitive Initiative: Leading the World in Innovation*. Washington, D.C.: Office of Science and Technology Policy.

Feldman, Maryann., David Guston, Stephen Hilgartner, Rachelle Hollander, Sheila Slaughter. 2003. *Research Policy as an Agent of Change: Workshop Report*. Washington, D.C.: National Science Foundation.

Friedman, Milton., and Anna Schwartz. 1963. *A Monetary History of the United States, 1870-1960.* Princeton, New Jersey: Princeton University Press.

Gerstein, Dean R., R. Duncan Luce, Neil J. Smelser, and Sonja Sperlich. 1988. *The Behavioral and Social Sciences: Achievements and Opportunities*. National Academy Press: Washington, D.C.

Granato, Jim., and Frank Scioli. 2004. "Puzzles, Proverbs, and Omega Matrices: The Scientific and Social Consequences of the Empirical Implications of Theoretical Models (EITM)." *Perspectives on Politics* 2: 313-323.

Granato, Jim., and M. C. Sunny Wong. 2006. *The Role of Policymakers in Business Cycle Fluctuations*. New York: Cambridge University Press.

Greenspan, Alan. 2004. "Risk and Uncertainty in Monetary Policy." Speech presented at the American Economic Association Annual Meeting, San Diego, California.

Haavelmo, Trygve. 1944. "The Probability Approach in Econometrics." Supplement to *Econometrica* 12: S1-115.

Heckman, James. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." Quarterly Journal of Economics 115: 45-97.

Kuhn, Thomas. 1969. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lucas, Robert. E., Jr. 1980. "Methods and Problems in Business Cycle Theory." *Journal of Money, Credit, and Banking* 12: 696-715.

Marburger III, John H. 2005. "Wanted: Better Benchmarks." *Science* 308, May 20: 1087.

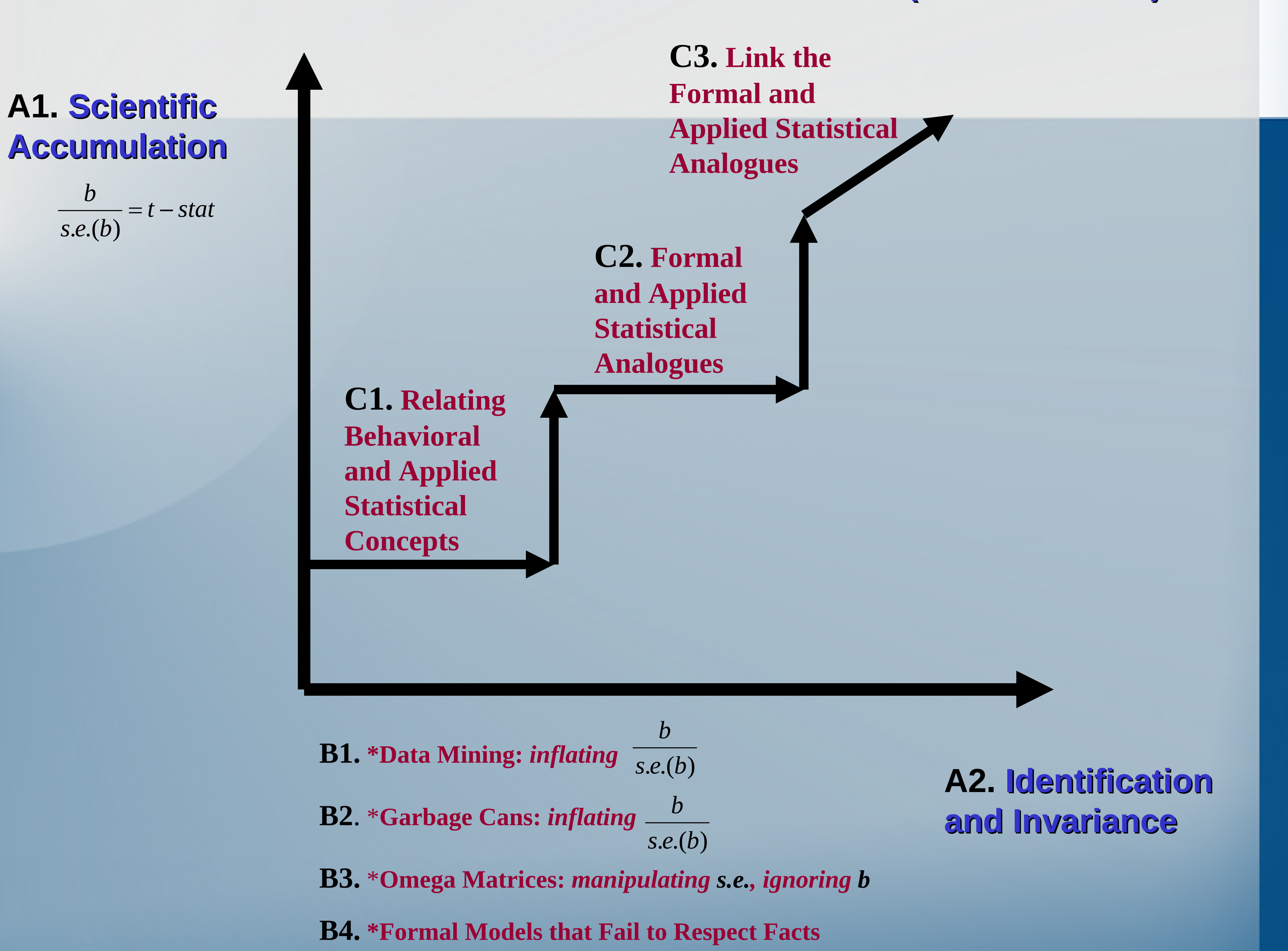
Meyer, Laurence. H. 2004. *A Term at the Fed: An Insider's View*. New York: HarperCollins.

Morgan, Mary. 1990. *The History of Econometric Ideas*. New York: Cambridge University Press.

Patton, Michael Q. 2002. *Qualitative Research and Evaluation Methods,* Third Edition. Thousand Oaks, California: Sage Publications.

Rossi, Peter., Mark W. Lipsey, and Howard E. Freeman. 2004. *Evaluation: A Systematic Approach,* Seventh Edition.

Figure 1


A3. Predictive Accuracy (Reassessment)
A1. Scientific Accumulation
$\frac{b}{s.e.(b)} = t - stat$
C3. Link the Formal and Applied Statistical Analogues
C2. Formal and Applied Statistical Analogues
C1. Relating Behavioral and Applied Statistical Concepts
B1. *Data Mining: *inflating* $\frac{b}{s.e.(b)}$
B2. *Garbage Cans: *inflating* $\frac{b}{s.e.(b)}$
B3. *Omega Matrices: *manipulating* **s.e.**, *ignoring* ***b***
B4. *Formal Models that Fail to Respect Facts
A2. Identification and Invariance