1

Not All Fun and Games

Challenges in Mathematical Modeling

INTRODUCTION

In large part, the inspiration for this book came from three sources, which can be categorized neatly as a failure, a challenge, and an ideal. First, the failure. When I began teaching in the profession, I was immediately assigned to graduate methods coursework. This is the experience of many professors trained in the last decade with a mathematical bent, and I was lucky enough to teach at an institution with an excellent culture. Unlike many other political science departments that exist in a state in which "there is war of every one against every one," Duke's political science department is almost entirely free of disputes about the value of mathematical modeling in the social sciences. Divisions of opinion certainly exist but, more or less, everyone in the department recognizes the virtue of mathematical methods for at least some problems.

Better still, even those who do not practice mathematical modeling believe in good research design. As many prospective faculty members discover during their job talks, "methods questions" and questions about research design are just as likely to come from the theorists of the department as anyone else (though couched in different terminology). Between job talks, faculty brown bags, and informal interactions graduate students have with faculty, it would be hard to finish a Ph.D. at Duke and not try your hand at mathematical modeling.

Despite this positive culture, teaching graduate methods coursework has not been easy. As has been noted in numerous places, the

shock most politics students experience on entering graduate school is severe. They expect to talk shop, debate the issues, and deal with "big" questions about the state of the world; instead, their first experience of graduate training at Duke involves a mathematics camp in the dull heat of August. No weighty matters of politics are discussed in this camp, unless one thinks that urns and the different colors of balls one places in them are of great import. Some students take years to get over this shock, essentially repeating much of their methods coursework when they come to a point in their own research where they have a pressing need for it. Others acquire good technical skills but nonetheless have great difficulty finding interesting questions or arriving at "good" models. Clearly, my best efforts were not sufficient and it drove me to think about issues of modeling in the social sciences and how one should attempt to improve matters.

In particular, why were so many bright graduate students, many of whom had good technical skills, unable to make the leap to generating testable theories? Why did many graduate students identify themselves primarily by their choice of method (e.g., game theory) rather than their research question? And, finally, were there any features of mathematical methodology in political science that added to the difficulty of training graduate students? These questions form a thread that continues throughout this book, and, hopefully, the questions offered here will demonstrate that many of the problems in training are related to conceptual problems in our mathematical methodologies.

The second influence on this book concerns a challenge to the discipline raised by Beck, King, and Zheng (hereafter, BKZ). Their paper – "Improving Quantitative Studies of International Conflict: A Conjecture" – appeared in the *American Political Science Review* in 2000. The paper was a broad challenge to empirical work throughout the social sciences, not just in international relations, and turned on the idea of what the proper relationship was between deductive models (usually represented by game theory) and empirical work (applied statistics). Normally, the ideal paper for the mathematical modeling crowd is a well-specified game that reaches some equilibrium outcome, which is then instantiated and tested in an appropriate statistical model. If hiring is any signal of departmental preferences, empirical work or game theoretic work alone is not as desirable as a combination of the two.

The importance of the paper by BKZ is that they argue for an entirely different approach. Instead of modeling the data generating process (DGP), they assume it is complex and interactive, and that prior efforts to model the origin of conflicts using game theory have not amounted to much (at least not anything testable). They conclude that the only reasonable standard for evaluating a statistical model is out-of-sample performance, without regard to the assumptions or specification of the statistical model. Not surprisingly, they adopt a nonparametric approach and use a neural network to generate an empirical model of conflict without regard to any underlying theory. Their article thus challenges the current methodological orientation of the discipline, insofar as they eschew the ideal of mapping a strategic game to an empirical specification.

I was confident that BKZ were wrong on several particulars, most notably whether their model actually outperformed the standard logit model used by many scholars in quantitative international relations. Along with Christopher Gelpi and Jeffrey Grynaviski, I wrote a reply addressing this problem. Additionally, we presented a general framework for comparing models when the goal is maximizing out-of-sample performance.[1] The larger epistemological questions raised by BKZ remained, however, and their challenge cast into doubt the proper relationship between deductive and empirical models. This dispute and how it relates to the broader themes of this book are dealt with starting in the next chapter.

The final source of inspiration that led to this book concerned an ideal of the proper approach to mathematical modeling in the social sciences. This ideal was first advanced in a set of workshops dubbed "Empirical Implications of Theoretical Models" (hereafter, EITM) funded by the National Science Foundation (NSF) in 2002. After these initial meetings, EITM evolved into a joint effort of Harvard, Michigan, Duke, and Berkeley to train advanced graduate students during the summer. Unlike other methods workshops that focus on particular skills (e.g., the Interuniversity Consortium for Political and Social Research's summer courses), EITM has the larger, epistemological goal of helping young researchers to bridge the divide between

[1] Our article, plus a response from BKZ, is in the May 2004 issue of the *American Political Science Review*.

deductive and empirical methodology. The goals of EITM were summarized in a 2002 report presented at the NSF:

> Significant scientific progress can be made by a synthesis of formal and empirical modeling. The advancement of this synthesis requires the highest possible levels of communication between the two groups. Formal modelers must subject their theories to closely related tests while, at the same time, empirical modelers must formalize their models before they conduct various statistical tests. The point is not to sacrifice logically coherent and mathematical models. Rather, it is to apply that same rigor to include new developments in bounded rationality, learning, and evolutionary modeling. These breakthroughs in theory will be accomplished with the assistance of empirical models in experimental and non-experimental settings.
>
> How will progress be measured? There are several performance indicators, including the number of articles that use formal and empirical analysis in the major professional journals. Another measurable indicator is the number of NSF grant proposal submissions by faculty and graduate students (doctoral dissertations) that use both approaches. However, the one area that may be the most difficult to measure is improvement in the quality of knowledge. In this regard, the ramifications of merging formal and empirical analysis is a transformation of how researchers think about problems and whether they take intellectual risks in synthesizing the model and testing it. When they do, the primary achievement of EITM will be a better understanding of the political and social world, more accurate predictions, and ultimately the provision of solid information to policymakers whose choices can profoundly affect citizens' quality of life.

Although out-of-sample forecasting is specifically emphasized in the above passage, it is obvious that the EITM founders have in mind something quite different than the nonparametric work of BKZ. Their goal is to rework the discipline so that the chasm between formal modelers and empirical researchers is bridged, with the hopes that this synthesis will lead to better models that have clearly testable empirical hypotheses.

By and large, I was very sympathetic to the goals of EITM, and was lucky enough to be invited to participate as a faculty member in the 2003 session at Michigan. My job seemed easy: take two days and present a framework for accomplishing EITM-style research. In my mind, this meant making an argument for how one might bridge the gap between models (usually deductive) and empirical tests; currently, the clearest statement of the difficulties inherent in this problem is

found in two articles by Signorino (1999) and Ramsay and Signorino (2003). After a bit of reflection, the issues involved were more difficult than I at first realized. Many of the arguments presented in this text are a direct result of the questions I faced in formulating my talks for EITM. Chapter 3 lays the groundwork for this investigation, and Chapters 4 and 5 provide a set of tentative answers to how one might implement the EITM statement on methodology.

WHAT THIS BOOK IS NOT

Before proceeding, it is important to say what this book is not. This book, despite appearances in some places, is not a critique of game theory (or formal theory more broadly). Although I am critical of some current practices, it should be obvious that I firmly believe in the aspirations of those who wish to make political science an actual science, complete with predictions and policy advice about events in the real world. My main concern is that game theory has become confused with definitions of human rationality. In this text, I will argue that game theory is a mathematical tool, not a proxy for human rationality where if one departs from game theoretic models one automatically sacrifices any notion of rational agents. As a tool, it is one way to "solve" problems and is better suited to some classes of problems than others. Most of the examples I focus upon concern classes of problems that for a number of reasons are ill-suited for a game theoretic approach, and I propose a set of methods "rational" agents might employ to deal with these complications. The reason for providing tools that expand the class of problems one can deal with analytically is in my mind simple: better models, with more verisimilitude, allow an easier transition to empirical tests. This is the primary goal advanced by EITM.

Game theory also has been confused with pure mathematics, insofar as many practitioners feel no need to connect their models to empirical tests. Much that masquerades under the classification of "theory building" is not worth the appendices, and one should question the usefulness of models that rely upon limiting assumptions to produce whatever narrow result is desired by the researcher.[2] Following Granger (1999),

[2] Arrow's impossibility theorem, in contrast, depends upon assumptions that are of substantive interest and produces a result that is extraordinarily broad.

the viewpoint adopted here is that the connection between theoretical models and their empirical referents needs to be direct enough such that we can be satisfied that the tests we conduct are actually dispositive. Dispositive tests distinguish the actual model (or data generating process) from the universe of possible models. This viewpoint is by no means new; rather, it has been the subject of debate in economics for decades.[3] What is perhaps new will be the particular modeling approach adopted here, which combines traditional game theoretic investigations with computational models. The reason for this union hopefully will become clear in subsequent chapters.

This book is also not a critique of empirical work in political science, though, again, one might be confused given that in places I am critical of existing efforts. Just as models without empirical tests are suspect, so, too, are data-driven statistical investigations that fail to make apparent what model is being tested. Good statistical work allows us to distinguish useful models from the universe of irrelevant models; further, it allows us to investigate the generality of a model and the places where assumptions are carrying too much of the load. I will, however, place rather more emphasis on predictive work than is currently the norm within the social sciences, as much of the statistical research that has been conducted in the social sciences aims solely at comparing the in-sample performance (or "explanatory power") of various models.

In-sample comparisons should be seen as innately suspect, as one can easily overfit a statistical model and claim "success" for a theory. More time will thus be spent in this text addressing the curse of dimensionality that has to this point been largely ignored by social scientists.

A Simple Example: Applause, Applause

As is fitting for a book on modeling, let us begin with a simple question. Hopefully, this will introduce most of my essential arguments before we wade into the deep end of the book. The history of this example is

[3] See, for example, the October 1993 *Special Issue Anniversary of the American Journal of Agricultural Economics*. Castle (1993) and Leontief (1993) are particularly useful in this issue, insofar as they outline a set of requirements that would help connect deductive models with empirical tests.

a rich one, given that it was used for many years by John Miller and Scott Page at their Computational Economics Workshop at the Santa Fe Institute.[4]

Imagine you are asked to explain or predict the occurrence of standing ovations. You have a performance of some type, where each member of the audience receives a signal from the performance about how good it is (based upon their own internal preferences). Each audience member can then choose to do nothing, applaud, or stand and applaud. They also can sit down again at any point should they decide to stand initially. This is a highly stylized problem but has relevance for social scientists. We often want to understand who stands, or votes, or participates in a riot, and how individual characteristics and social dynamics lead to this behavior.

There are different approaches one might take to this problem, and in social science one can roughly describe the three methodological traditions that could be utilized: empirical, deductive (i.e., game theoretic), and computational. Let us investigate what sorts of answers these traditions, in isolation, might provide to the standing ovation problem.

An empirical researcher would likely start out with questions concerning what measures would be collected for both the dependent and independent variables, and not all of the forms of these measures would be obvious. For example, the dependent variable might be coded as a binary variable measuring whether or not the ovation occurred. If this encoding is adopted, what would the right threshold be for distinguishing an ovation? Would 90% have to stand? More? Less? The choice of scale for the dependent variable is also not obvious; one could change both the temporal and spatial characteristics of the dependent variable. For example, one encoding would measure the length in time of the ovation, but any such measure of time would retain the problem of choosing an appropriate threshold. Alternately, one could measure the likelihood that any given audience member participates in the ovation, thus changing the unit of analysis spatially from the entire audience to each individual member.

[4] Past answers to this problem are archived at http://zia.hss.cmu.edu/econ/homework95.html. For the most recent investigation of this problem, see Miller and Page (forthcoming).

Another more insidious problem would involve the nonindependence of observations.[5] Cleary, if subsets of the data set involved repeat performances by the same artist, "buzz" might result in a lack of independent, identically distributed (IID) observations. This problem also would complicate the measure of independent variables. Measures of performance quality and the like could easily be contaminated by interactions either between guests of the same performance (e.g., social pressure) or for members that attend multiple performances across observations in the data set. And members of different audiences are obviously not drawn from identical distributions, as people sometimes choose which performances they attend.

Problems aside, what sorts of questions would the empirical researcher answer? Likely, it would involve establishing relationships between such concepts as "performance quality" (as perceived by the audience), the type of performance, the number of audience members, and so on, and the likelihood or length of an ovation.

A deductive (or formal) modeler would come at this problem from a different angle, where the most important decision would involve specifying the benefits and costs that are present for members of the audience when they decide to ovate or not. Clearly, you do not want to be the only fool in the audience standing and clapping madly; people would stare. Just as clear, you do not want to be the grinch, sitting alone in a sea of excited fans. At some level, though "quality" matters, you only want to reward "good" performances with an ovation, given the effort involved in standing and clapping.

The structure of the game would also involve a set of important considerations on the part of the deductive modeler. How many periods would be included in the game where agents could update their information and make choices? If an ovation occurred, how would people get back to their seats? The same sorts of utility considerations discussed in the preceding paragraph would apply with equal force to agents making choices to sit back down again.

Given these modeling choices, and the input of a few "state of nature variables" such as the quality of the performance, the deductive

[5] One also might point out that the observations are not independent spatially – that is, whether or not one member of the audience stands (or later, sits) is likely correlated with the actions of other audience members.

modeler might well reach a good understanding of the individual decisions that work together to produce an ovation. A model might also help worried performance-goers in reaching decisions about whether or not to stand for an ovation in future performances. Ever present, however, would be the worry that the limiting assumptions relied upon to formulate a sufficiently simple model might cut against the usefulness of any insight gained.

The final tradition that might generate a solution for this problem is less well known in the social sciences. A computational (or dynamic systems) researcher, in contrast to the two preceding approaches, would specify a set of rules that governed the behavior of individual audience members, along with a set of contextual variables that described such features as the seating arrangement, the shape of the performance hall, relationships between audience members, and so on. What would these rules look like? On one level, the rules would be functional expressions that would be similar to the utility functions used by a game theorist, though these functions might well be allowed to vary both in time and by the individual type of audience member. On another level, these rules could add substantial verisimilitude to the computational model by incorporating features of the problem that would be difficult to model in a deductive framework (e.g., learning models based upon research in cognitive psychology). One such rule might involve adding vision to the model – given the shape of the performance hall, not all audience members can physically see all other audience members. Any utility function that involved peer pressure should be more sensitive to people within an agent's field of vision than agents outside this field.[6]

Unlike a game theoretic model, it is unlikely that a computational model would produce a set of deductive results. What is far more likely is that the researcher, confronted with the large parameter space generated by the rules used in formulating the computational model, would have to rely upon statistical investigations to understand

[6] The outcome of such a rule is that not all audience members are created equal – that is, audience members in the middle rows nearest the stage would have a disproportionate share of influence. One also might consider the type of individual audience members. For example, if a group of Catholics got together to watch a play, it might matter if the Pope were sitting in the audience. I would hazard that if the Pope ovates, so, too, would everyone else.

any "results" of the computational model, much as in the empirical tradition. Statistical relationships between parameters, rules of interest, and the likelihood of an ovation would then be presented, albeit substituting artificial data for real data.

This is a brief sketch of an interesting problem, but it raises questions of importance to all modelers. To begin with, are the approaches complementary or distinct? On the face of it, our three stereotypical methodologists would not have much to say to each other. The empirical researcher is establishing correlations between different measures and the likelihood of ovation; the game theorist provides advice on how rational audience members should select strategies; and the computational modeler incorporates aspects of both of the forgoing approaches to produce a dynamic model that recreates a standing ovation.

All of these models ostensibly explain the same phenomenon, but can one compare or integrate the results? Or, are these simply different answers to different questions? I will argue in the succeeding chapters that it is undesirable to let each type of modeler work in a vacuum; models need to produce results that are directly comparable to competing explanations. Even within each methodological approach, models are often not unique. Different modelers will produce different answers, and the job of social science should be to sort among them by insisting on out-of-sample tests of some kind. If, for example, we are confronted with several different game theoretic models, all explaining standing ovations, how do we decide which one is closest to being right? Unless one of the game theorists makes a deductive mistake, the models will differ because the assumptions differ. Arguing about assumptions is a little like arguing about whether Wolverine is tougher than the Hulk; ultimately, it comes down to taste. This book will argue that a different, integrated approach is required to make sense of these questions.

STRIFE BETWEEN METHODOLOGICAL CAMPS

Currently, there is a sense of mutual distrust between different methodological camps. Let us start with the more forceful critiques of the empirical tradition. As part of the EITM meetings, Christopher Achen argued that one must be suspicious of empirical modeling in the social sciences. Because many models are quite complex, researchers have

an abundance of parameter choices that allow them to overfit models, generating any outcome they wish:

> Empirical work, the way too many political scientists do it, is relatively easy. Gather the data, run the regression/MLE with the usual list of control variables, report the significance tests, and announce that one's pet variable "passed." This dreary hypothesis-testing framework is sometimes even insisted upon by journal editors. Being purely mechanical, it saves a great deal of thinking and anxiety, and cannot help being popular. But obviously, it has to go. Our best empirical generalizations do not derive from that kind of work. How to stop it? The key point is that no one can know whether regressions and MLEs actually fit the data when there are more than two or three independent variables. These high-dimensional explanatory spaces will wrap themselves around any data set, but typically by distorting what is going on. They find the crudest correlations of course: education increases support for abortion, for example. In the behavioral tradition, that counts as a reliable finding. But no one knows why education is associated with that moral position (higher intellect discovering the truth? mindless adoption of elite tribal norms? correlation with something else entirely?), and that leaves open the possibility that abortion attitudes do not work the way the literature says they do. Getting rid of this cheap sense of "empirical findings" is probably the central task that empirical political research faces. . . .
>
> As an instance of the altered perspective I have in mind, I propose the following simple rule: Any statistical specification with more than three independent variables should be disregarded as meaningless. With more variables than that, no one can do the careful data analysis to be sure that the model specification is what s/he says it is. (Achen in the National Science Foundation EITM Report, 2002, Appendix B)

Or, one might look farther back to Keynes, and his critique of the hapless Professor Tinbergen:

> I infer that he considers independence of no importance. But my mind goes back to the days when Mr. Yule sprang a mine under the contraptions of optimistic statisticians by his discovery of spurious correlation. In plain terms, it is evident that if what is really the same factor is appearing in several places under various disguises, a free choice of regression coefficients can lead to strange results. It becomes like those puzzles for children where you write down your age, multiply, add this and that, subtract something else, and eventually end up with the number of the Beast in Revelation. . . .
>
> To the best of my understanding, Prof. Tinbergen is not presented with his time-lags, as he is with his qualitative analysis, by his economist friends, but invents them for himself. This he seems to do by some sort of trial-and-error

method. That is to say, he fidgets about until he finds a time-lag which does not fit in too badly with the theory he is testing and with the general presuppositions of his method. No example is given of the process of determining time-lags which appear, when the come, ready-made. But, there is another passage where Prof. Tinbergen seems to agree that the time-lags must be given a priori. . . .

These many doubts are superimposed on the frightful inadequacy of most of the statistics employed, a difficulty so obvious and so inevitable that it is scarcely worth the time to dwell on it. (Keynes 1939)

At root, Achen and Keynes are addressing the same problem in empirical methods. Unbeknown to anyone save the original researcher, choices are made in empirical work. *Lots* of choices. Given the obvious problem of false correlation, it does not seem too much of a stretch to imagine that any empirical modeler, given time, can produce almost any result that is desired. Journals and monographs, by their nature, only report "positive" results and only the "final" model. How much pain or guesswork or outright cheating at the margins that goes into an empirical paper is never seen in print. One way to think of this is to imagine every salient choice made by the empirical modeler as a parameter; results are thus conditional statements made upon the particular set of parameter values chosen. Given how large these implied parameter spaces are, one cannot place much faith in a final report of in-sample performance.

A deductive modeler (typically relying upon game theory) would certainly agree with the forgoing critique of statistical methods/econometrics. Moreover, most formal theorists believe that their methodological approach is immune to the flaws that plague other approaches. Niou and Ordeshook, for example, cite the transparency of formal theory as an enormous advantage over both qualitative and empirical methodologies:

But the rational choice paradigm and formalism are not mushrooms that sprung up in an unattended intellectual forest. They are reactions to a discipline mired in imprecision, vagueness, obscure logic, ill-defined constructs, nontestable hypotheses, and ad hoc argument. They are a reaction to a discipline that in the 1920s proclaimed the Weimar constitution the greatest political-intellectual achievement of its age; a discipline that in the 1960s substituted correlation for cause; a discipline submerged in such conveniently vague and ill-defined ideas as "power," "leadership," "authority," "group," "alliance," "function," "ideology," "culture," "regime," "stability," and "balance." They

are reactions to a discipline that substituted the well-turned phrase for concrete constructs, operational measures for theoretical primitives, and the gloss of methodological sophistication for true theory. They are, in short, a reaction to a discipline that did and does precisely what Walt critiques the formal analyst of doing – burying key assumptions in an indecipherable format, although generally that format was a language more to the liking of those who studied French and Plato in college rather than calculus. (Niou and Ordeshook 1999, 87)

In addition, it is obvious that Niou and Ordeshook draw a sharp distinction between the results of formal theory (which are uncontestable; that is, the results follow deductively from the premises) and empirical work that could be rife with spurious correlation. Bueno de Mesquita and Morrow go even further in a defense of formal theory by arguing that of all the virtues one might discover in a social science theory, logical consistency is foremost:

Walt gives three criteria for evaluating social science theories: logical consistency, degree of originality, and empirical validity. We believe that logical consistency takes precedence over the other two criteria; without logical consistency, neither the originality of a theory nor its empirical validity can be judged. Logical consistency is the first test of a theory because consistency is necessary, though not sufficient, for understanding how international politics works.

A basic point in logic drives our view. A theory, in terms of logic, consists of a system of assumptions and conclusions derived from those assumptions. A logical inconsistency exists when two mutually contradictory statements can be derived from the assumptions of a theory. When such a contradiction exists in a theory, then any statement follows logically from the theory. There is, then, no discipline for arguments in a logically inconsistent theory; those using the theory are free to draw any conclusion they wish from the premises of the theory. Logical inconsistencies deny the possibility of a theory having empirical content. Theories derive empirical content by producing falsifiable hypotheses, conclusions that could be contradicted by evidence. A theory gains credence as more of its falsifiable propositions are supported by evidence, although there are no hard and fast rules here. However, because any pattern of evidence can be matched with some conclusion of a logically inconsistent theory, such theories cannot be falsified and so cannot have empirical content. A theory is falsified when an alternative is shown to fit the range of predictions better than the initial theory. Falsification of a theory cannot happen if any evidence can be interpreted as an implication of the theory. . . .

Again, any conclusion can be derived when a logical inconsistency exists, and so the choice of which conclusion to use for policy purposes falls entirely on the tastes or prejudices of the party making the prescription. Indeed, the use

of a logically inconsistent theory to justify a policy recommendation is worse than recommendations not supported by any theory. . . .

For these reasons, we believe that logical consistency has pride of place among the criteria for judging social science theories. (Bueno de Mesquita and Morrow 1999, 56–7)

There is, of course, a problem with arguments that attempt to draw a sharp distinction between empirical modeling and formal theory. Although it is the case, as Bueno de Mesquita and Morrow note, that "any pattern of evidence can be matched with some conclusion of a logically inconsistent theory," the exact same statement is true of logically *consistent* theory. As should be obvious (but for some reason is not), one may achieve any outcome one desires with consistent theory; all it takes is the right combination of assumptions, solution concepts, and the like. The chore for formal theory cannot rest solely upon consistency, as *the class of "consistent" games that provide any given result is infinite*.

One can think of this argument in a different way. Imagine a researcher perceives an empirical regularity – for example, that candidates tend to take positions near the middle of a left-right ideological dimension. How many consistent models could the researcher construct that would produce center-seeking candidates? Infinitely many. And only some of the models are "right" in the sense that they are analogous to the real, underlying process. All the other (infinite) models are correlated with the empirical regularity in much the same way that an empirical specification is spuriously correlated with a given sample. Without finding novel data or deriving secondary conclusions, one cannot place much certainty in any single model of center-seeking candidates.[7]

So while one must agree with Bueno de Mesquita and Morrow that consistency is a necessary condition, the more important goal is to choose the "best" theory from the class of consistent theories that produce a desired result. Friedman (1953) sums up the problem created by attributing virtue to consistency alone:

Logical completeness and consistency are relevant but play a subsidiary role; their function is to assure that the hypothesis says what it is intended to say

[7] For a similar perspective worth further study, see Lave and March (1975).

and does so alike for all users – they play the same role here as checks for arithmetical accuracy do in statistical computations. One effect of the difficulty of testing substantive economic hypotheses has been to foster a retreat into purely formal or tautological analysis. As already, noted, tautologies have an extremely important place in economics and other sciences as a specialized language or "analytical filing system." Beyond this, formal logic and mathematics, which are both tautologies, are essential aids in checking the correctness of reasoning, discovering the implications of hypotheses, and determining whether supposedly different hypotheses may not really be equivalent or wherein the differences lie. . . . But economic theory must be more than a structure of tautologies if it is to be able to predict and not merely describe the consequences of action; if it is to be something different from disguised mathematics. And the usefulness of the tautologies themselves ultimately depends, as noted above, on the acceptability of the substantive hypotheses that suggest the particular categories into which they organize the refractory empirical phenomena. (Friedman 1953, 11–12)

Choosing a game that provides a given result (that you want to achieve *a priori*) is thus not at all different than the problem of false correlation in the statistical literature. Not only is this always possible, it is also the case that the mapping of formal theories to results is not a one-to-one correspondence.[8] One might appeal to maxims such as parsimony, or generalizability (or whatever) to discriminate between competing formal theories, but this is very slippery epistemological ground, and places such discrimination firmly in the land of taste rather than science.

Moreover, all choices that go into a particular formal theory that are left to the modeler should be seen as traversing a very large parameter space; again, this problem mirrors the corresponding complaint levied against empirical modelers. As Peltzman (1991) notes, "Game theory has introduced a rigor in the analysis of rational behavior that was missing [but] skepticism about the marginal value of recent theory is warranted [because] conclusions drawn tend to be very sensitive to the way problems are defined and to the assumptions that follow." Game theoretic results are conditional upon these choices, and given

[8] Other than trivial examples, it is clear that the mapping of theories X to results Y is a bijection but not an injection. Rather, mapping theories to results is a many-to-one process, and the goal of formal theory should be to sort between the class of possible theories.

the size of these parameter spaces, results must be seen as exceptionally brittle things when the only test is whether or not the formal theory produces an expected outcome. The problems that plague empirical methodologies thus have almost perfect analogues in formal modeling.

One example of the forgoing pathology in formal theory can be found in the dispute between Banks (2000) and Groseclose and Snyder (2000) in the pages of the *American Political Science Review*. Banks, in short, shows that one of the results in Groseclose and Snyder's original paper on creating supermajorities in legislatures is wrong; that is, it fails the consistency condition raised above. The response, however, by Groseclose and Snyder is illuminating, insofar as they simply change an assumption such that the original result holds. As they note, the changed assumption "is crucial for our [Groseclose and Snyder's] results" and "the opposite assumption is crucial for Banks's results" (Groseclose and Snyder 2000, 683).[9] If one perceives that minimal winning coalitions are rare in actual legislatures, this dispute proves that one can certainly arrive at a model that yields that general result, even if one stumbles along the way. Further, it shows that although game theoretic results are in principle transparent, this is not necessarily the case in practice. The Groseclose and Snyder result was in print for four years before the error was found.[10]

Does computational modeling have similar defects? Of course it does. Like the formal theorists, computational modelers often claim that they also have transparent models. Instead of presenting a list of assumptions as a *fait accompli* as formal theorists do, the best computational models typically provide not only the assumptions but also an idea of what happens to the model's results when the assumptions are modified. But, despite this potential advantage, the fact remains that most social scientists cannot be expected to wade through thousands of lines of C++ code to understand the inner workings of a computational model, nor do journals and book editors publish such details. Just as

[9] As Groseclose notes (personal communication), empirical work could in principle distinguish between these competing models. Yet, the articles in this debate are entirely absent empirical work, which forces one to argue about assumptions rather than the question at hand (that is, the actual frequency of minimal winning coalitions).

[10] In addition, the normal review process would indicate that three referees also missed the error, as did the dozen or so citations of the article that occurred before Banks's reply (according to the ISS Web of Science).

with empirical and formal models, we are left with a situation in which one can write a computational model (actually, infinitely many) that will (with the right parameter settings, rules, etc.) produce any given result.

The only qualitative differences between computational models and formal theory is that computational models are rather more ecumenical in how they encode problems. Additionally, computational models often possess more verisimilitude at the cost of deductive tractability. One does not "solve" a computational model; one uses it to generate simulated data that one tests with the tools of applied statistics. Computational models are thus related to game theoretic models, except that they usually address more complex problems and lack deductive solutions (but, more on this in Chapter 3).

What sorts of additional problems plague computational models? Take for example the outputs of three models captured in Figure 1.1. Absent any additional information, it is difficult to discern what these slides are showing. All three look very much alike, though there are some differences in the level of clustering apparent in the slides. It may come as a surprise, then, that each of these slides purports to demonstrate a different computational "result," explaining such diverse phenomena as state formation (Cederman 1994), culture dissemination

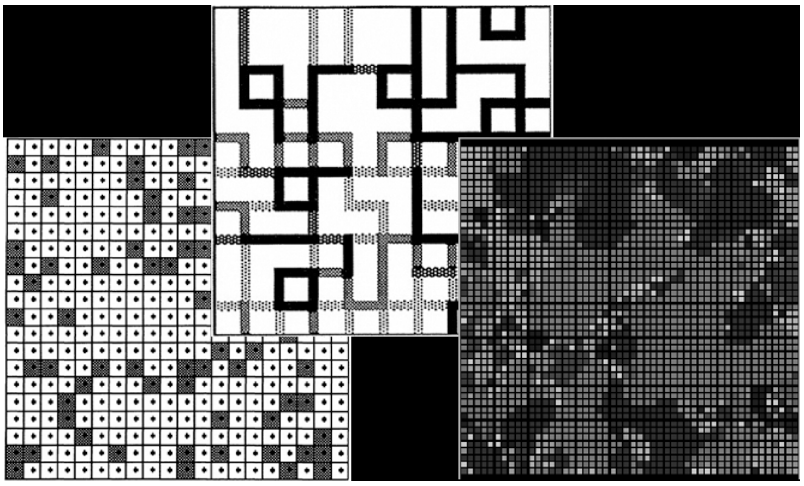


Figure 1.1. An Example from Computational Modeling

(Axelrod 1997), and collective identity (Lustick 1999).[11] While it is the case that the underpinnings of all of the models are related, this is not a cause for rejoicing – unless one has hopes that social science has stumbled on another iterated prisoner's dilemma. Each model is based upon an Ising model borrowed from physics, where it is used in statistical mechanics. By tweaking parameters (e.g., whether the neighborhood is Von Neumann or Moore), each author produced qualitative output that for whatever reason was suggestive to them. Given how slippery evaluations can be of such visual output, plus the huge parameter space underlying all such models, it is difficult to be any more optimistic about these results than those from any other methodological approach.

The main question is what one does about these difficulties. As we have seen, the problems that haunt the various methodological schools are more similar than they first appear, and the main goal of this book is to propose a solution to these problems. The short answer is that a structured combination of the methodological approaches I have listed is far superior to any approach taken separately. Much of the rest of this book will be spent examining what this "combination" looks like. While this position is at present only sketched, a brief tour of epistemology will help motivate the more detailed proposals that follow.

A SHORT STATEMENT ON EPISTEMOLOGY

From the preceding discussion, I have sketched a few of the problems that complicate the use of mathematical methods in the social sciences. If things are to improve, I would argue that a shift in our underlying epistemology is needed. The argument presented here is very close to the classic statement of Friedman (1953), and it is worth exploring how Friedman's view of epistemology has been critiqued

[11] It is important to note that Lustick's model is available for download in a format in which one can easily modify parameter values to test their impact on the results. This is extraordinarily helpful, but falls short of best practice, insofar as what computational modeling needs is a clear result that provides better predictive leverage on a question researchers care about. Like game theory, computational models most often serve as existence proofs.

(and subsequently ignored). As noted by Hausman, Friedman is not a standard instrumentalist:

> Friedman declares, "The ultimate goal of a positive science is the development of a 'theory' or 'hypothesis' that yields valid and meaningful (i.e., not truistic) predictions about phenomena not yet observed" (p. 7). This is the central thesis of instrumentalism. But from a standard instrumentalist perspective, in which all the observable consequences of a theory are significant, it is impossible to defend Friedman's central claim that the realism of assumptions is irrelevant to the assessment of a scientific theory. For the assumptions of economics are testable, and a standard instrumentalist would not dismiss apparent disconfirmations. (Hausman 1984, 217)

What troubles Hausman about Friedman's modification of instrumentalism? In Friedman's words,

> Viewed as a body of substantive hypotheses, theory is to be judged by its predictive power for the *class of phenomena which it is intended to "explain"*. Only factual evidence can show whether it is "right" or "wrong" or, better, tentatively "accepted" as valid or "rejected." As I shall argue at greater length below, the only relevant test of the validity of a hypothesis is comparison of its predictions with experience. The hypothesis is rejected if its predictions are contradicted ("frequently" or more often than predictions from an alternative hypothesis); it is accepted if its predictions are not contradicted; great confidence is attached to it if it has survived many opportunities for contradiction. Factual evidence can never "prove" a hypothesis; it can only fail to disprove it, which is what we generally mean when we say, somewhat inexactly, that the hypothesis has been "confirmed" by experience. (Friedman 1953, 8–9; emphasis added)

The distinction between proving a theory false and confirming a theory by experience is nothing new; most texts on modeling in the social sciences have adopted some version of Popper's work on falsification. What is new is that Freidman, as Hausman points out, limits the investigation or testing of a theory to the particular dependent variable the theory aims to explain. Under this limitation, attacking a rational choice model by "proving" the assumptions are not held by actual human actors is entirely beside the point. Experiments of the sort conducted by Kahneman and Tversky (1979) are useless in critiquing the results of a rational choice model designed to study a particular phenomenon; unless, of course, prospect theory has better

predictive ability for the phenomenon in question. Thus, in Friedman's terms, one can only compare theories based upon how well they predict out-of-sample, and the theorist is allowed to pick the dependent variable.

Hausman's problem with this statement is best revealed by his own example:

> I suggest that Friedman uses this view that science aims at narrow predictive success as a premise in the following implicit argument:
>
> 1. A good hypothesis provides valid and meaningful predictions concerning the class of phenomena it is intended to explain. (premise)
> 2. The only test of whether an hypothesis is a good hypothesis is whether it provides valid and meaningful predictions concerning the class of phenomena it is intended to explain. (invalidly from 1)
> 3. Any other facts about an hypothesis, including whether its assumptions are realistic, are irrelevant to its scientific assessment. (trivially from 2).
>
> If (1) the criterion of a good theory is narrow predictive success, then surely (2) the test of a good theory is narrow predictive success, and Friedman's claim that the realism of assumptions is irrelevant follows trivially. This is a tempting and persuasive argument.
>
> But it is fallacious. (2) is not true, and it does not follow from (1). To see why, consider the following analogous argument.
>
> 1′. A good used car drives safely, economically and comfortably. (oversimplified premise)
> 2′. The only test of whether a used car is a good used car is to check whether it drives safely, economically and comfortably. (invalidly from 1′)
> 3′. Anything one discovers by opening the hood and checking the separate components of a used car is irrelevant to its assessment. (trivially from 2′)
>
> Presumably nobody believes 3′. What is wrong with the argument? It assumes that a road test is a conclusive test of a car's future performance. (Hausman 1984, 218)

Hausman's example is quite nice: Assume one has a theory that predicts car performance (i.e., does the vehicle drive "safely, economically, and comfortably") based upon a test drive (in which "test drive" is the theory that produces an expectation about the car performance). Further assume that one can take the car to a mechanic, and that the

mechanic can open the hood and evaluate (in his mind) the status of various components of the car.

But Hausman makes a logical mistake in the above analysis. The main problem with Hausman's argument concerns his use of the term "only" both in propositions (2) and (2′). Proposition (2) is (in Friedman's terms) correct, but proposition (2′) is a misuse of Friedman and contains a logical contradiction. 2′ states that the only test of a used car is to see if it drives well, but Friedman certainly does not mean this. Many theories could be proposed other than a test drive to determine the quality of a used car; the only qualification Friedman raises is that all theories have the same empirical referent. If the mechanic in 3′ points at a component and states that a component is flawed, one has two choices. Either this theory (i.e., flaws in components imply poor car quality) has an implication for the used car's quality, or else it has no bearing at all on overall quality. If the former is true, Friedman places the implicit theory of 3′ on equal footing with the test drive theory; the way one chooses between the two theories is to examine their out-of-sample performance. If the latter is true – a mechanic inspecting components is unwilling to make a statement about the used car's quality – one must agree with Friedman in saying that this statement has little bearing on evaluating our "test drive" theory.

Proposition (2′) is thus false, but only because Hausman failed to map his proposition to what Friedman is actually saying. Friedman believes that there is a universe of models, not just one. Thus, the use of the term "only" in reference to the "test drive" theory in proposition (2′) is unwarranted:

> Additional evidence with which the hypothesis is to be consistent may rule out some of these possibilities; it can never reduce them to a single possibility alone capable of being consistent with the finite evidence. (Friedman 1953, 9–10)

Proposition (2′) and (3′), properly restated to be in accord with Friedman, should be:

> 2′. Testing a used car's quality by seeing if it drives safely, economically and comfortably is one theory of many. One selects from the universe of possible theories by relying upon out-of-sample performance. This selection is also supplemented by consideration of the "fruitfulness" and "simplicity" of a theory.

3′. Anything one discovers by opening the hood and checking the separate components of a used car is irrelevant to its assessment unless one develops a mapping between the mechanic's assessment and the used car's quality.

Of course, what Hausman means to say is that if the mechanic looks at the engine and sees something wrong, we know for a fact that the mechanic is "right" and the test drive is "wrong." But this kind of classification makes no deductive sense, especially if one has ever visited a mechanic or listened to Car Talk on National Public Radio. Real-world mechanics get things wrong all the time, and there is no reason whatsoever to privilege the "mechanic's evaluation" over the "test drive" theory. Friedman's claim that one should treat both of these as competing theories, and adjudicate between them based upon out-of-sample performance, is thus not only logically consistent but far superior to Hausman's classification, which depends upon an unstated and unsupportable leap of faith in the mechanic.

In all of the above, one must distinguish between deductive logic and probabilistic knowledge.[12] If p is a model and q an implication or test of that model, a restatement of Hausman's critique of Friedman seems to be:

i. $p \Leftrightarrow q$
ii. $\sim p \rightarrow \sim q$
iii. Show $\sim p$ to prove that the model is wrong.

Step iii. contains Hausman's argument in a nutshell: showing that p is false – either because the model is inconsistent or because the assumptions are wrong – is all that is required to reject the model. In particular, many within the social sciences advocate scrutinizing the assumptions of a model and are reluctant to accept models that depend upon assumptions known to be false.

This line of attack misses something fundamental about research, however. Models are probabilistic in nature and one often chooses to model a phenomenon at a tractable level of granularity given the precise question asked or the data that are available. Thus, models are

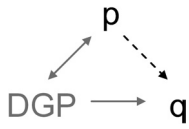
[12] I am obliged to John Aldrich for this example.

rarely unique, as the use of "if and only if" in (i) implies. As Nagel notes:

> In any event, physicists show no noticeable compunction in using one theory for dealing with one class of problems and an apparently discordant theory for handling another class . . . They introduce considerations based on the theory of relativity in applying quantum mechanics to the analysis of the fine structure of spectral lines; they ignore such considerations when quantum theory is exploited for analyzing the nature of chemical bonds. (Nagel 1961, 133–4)

When testing a used car, it is possible that several models all seek to explain the performance of the car and no dispositive test exists to sort between them. Better to adopt Friedman's famous "as if" approach to theories and allow for multiple theories than to decide that some sorts of knowledge (e.g., the mechanic) are privileged.[13]

A better approach than a purely deductive formulation is illustrated by the following graph:



There exists, we would hope, a "true" data generating process that produces the sample *q*; to our sadness, we will never know what this process is with any accuracy for most phenomena of interest. Our model *p*, however, may reflect something systematic about the DGP, though it is unlikely it will capture everything systematic about a complex process.[14] A model *p* that captures something (though not everything) essential of the DGP trumps other models, however, that fail to mirror

[13] Compare Friedman's "as if" argument to Nagel's description of physics:

> Everything depends on the problem; there is no inconsistency in regarding the same firm *as if* it were a perfect competitor for one problem, and a monopolist for another, just as there is none in regarding the same chalk mark as a Euclidean line for one problem, a Euclidean surface for a second, and a Euclidean solid for a third. (Friedman 1953, 36; emphasis added)

[14] This argument has caused a good deal of controversy – see footnote 14 in this chapter. As argued earlier, it is also likely that *p* is not unique – there are a multitude of models that all reflect different parts of the true DGP – the double arrow between *p* and the DGP thus reflects Friedman's controversial use of "as if" in his essay.

the DGP. Thus, it is not enough to say that if the assumptions of p are false then the model must be discarded.[15] In the case of statistical models, we accept arguments like this one without question and there is no reason the analogy does not hold for models more generally.

I would thus argue that Friedman's statement on epistemology is more compelling than critics have allowed. I would, however, make two amendments to Friedman's version of instrumentalism.

Amendment 1: Constraints upon Assumptions

On the one hand, I agree that assumptions are often proximate measures for more complicated phenomena, and to the extent they correlate with the real-world process, it is difficult to place much value in criticizing assumptions. As Friedman notes:

> Misunderstanding about this apparently straightforward process centers on the phrase "the class of phenomena the hypothesis is designed to explain." The difficulty in the social sciences of getting new evidence for this class of phenomena and of judging its conformity with the implications of the hypothesis makes it tempting to suppose that other, more readily available, evidence is equally relevant to the validity of the hypothesis-to suppose that hypotheses have not only "implications" but also "assumptions" and that the conformity of these "assumptions" to "reality" is a test of the validity of the hypothesis different from or additional to the test by implications. This widely held view is fundamentally wrong and productive of much mischief. . . . Truly important and significant hypotheses will be found to have "assumptions" that are wildly inaccurate descriptive representations of reality, and, in general, the more significant the theory, the more unrealistic the assumptions (in this sense). The reason is simple. A hypothesis is important if it "explains" much by little, that is, if it abstracts the common and crucial elements from the mass of complex and detailed circumstances surrounding the phenomena to be explained and permits valid predictions on the basis of them alone. (Friedman 1953, 14)

On the other hand, there is the problem of false correlation. One may always discover a model that predicts a given empirical referent, and it is difficult at times to know whether this discovery constitutes

[15] It is always possible that the nonsystematic component of q is what the model actually exploits. One can thus have the appearance of a good model when in fact p only predicts a component of q that is accidental. This is the main reason why one must always compare models based upon out-of-sample tests.

an advance of knowledge. An oft-used example is using the winning conference in the Super Bowl to predict the outcome of the presidential race – whatever theory that is advanced in defense of this result would be met with a great deal of skepticism, no matter how accurately it predicted the presidential race out-of-sample.

Hinich and Munger (1997), in their text on analytic methods, by and large adopt Friedman's perspective on epistemology, with one exception that addresses the forgoing concern about assumptions. They add the criterion that assumptions must be plausible, because opaque assumptions make it difficult to understand how brittle a model's result is:

> We claimed above that a strength of mathematical models is the clarity of the statement of the assumptions. Yet clarity is only a strength if the assumptions themselves are plausible. One cannot tell if an argument works outside its own stylized context by looking only at the argument itself. Consequently, the external application, or "testing," of formal theory is by analogy: The theory is tested by measuring relationships among observable phenomena, in hopes that the observable phenomena are "like" the relationships the model focuses on. Without careful empirical tests, models would just be amusing mathematical exercises. (Hinich and Munger 1997, 5)

Like Friedman, Hinich and Munger also note that empirical failure or falsification is one of the key motive forces in improving models. Why then do Hinich and Munger add the consideration of "plausibility" to Friedman's statement?

While Hinich and Munger do not precisely define a plausibility standard for assumptions, this kind of concept is echoed throughout much of the public choice school, and there is an expectation that assumptions have something to do with the phenomena under investigation. Mueller (2003) and Aldrich (1995, 1997), for example, discuss the idea of adding a constant to models of turnout. As most everyone learns in their first formal theory class, rational voters, knowing that the odds of affecting the outcome of an election (either because their vote decides the election outright or because their vote causes a tie) are negligible, cannot justify turning out to vote because of whatever benefits might accrue due to their preferred candidate winning election. Following Aldrich, if p represents the odds of affecting turnout, B represents the benefit derived from one's preferred candidate winning, c is the cost

of voting, and d is the intrinsic benefit of voting (e.g., the expression of citizenship), then when

$$pB - c + d > 0$$

a citizen will vote. Given that p is in most elections arbitrarily close to 0, one has to believe that $d > c$ to avoid universal abstention.

Most models avoid the result of zero turnout by theorizing about the role of d; Mueller points out that although adding the assumption that $d > c$ to models of turnout may seem plausible, it is very difficult to say what the assumption represents. One may claim it stands for civic duty, but just as easily someone else could say the constant stands for the utility of voting as an expressive act. A plausible assumption, then, for Mueller boils down to one's ability to map the assumption to the phenomena in question (i.e., as with Hinich and Munger, the assumption should relate to turnout and decision making). Additionally, the assumption should allow the researcher to distinguish between rival hypotheses.

Here I will settle on a distinction that is broad and builds upon Hinich and Munger's argument that assumptions are related to the phenomena in question and Mueller's additional constraint that assumptions uniquely identify concepts. Assumptions are plausible if three conditions are met:

I. The assumption is related to the phenomena under investigation in a fashion that is not absurd. That is, assumptions may well be gross simplifications of reality. As Friedman notes, the best models are simple, yet nonetheless provide great predictive leverage. "Assume a frictionless surface" is certainly false in many contexts, but it is simple to relate this assumption to models that study motion. As such, it should be preserved without criticism and one should follow Friedman's advice that the way to compare models is to compare out-of-sample performance on the dependent variable selected by the researcher. Arguments that a given model's predictive power is void because actual surfaces are not frictionless would *not* be compelling.
II. The assumption is organically related to logic of the model. For example, one might create a rational investment model that predicts little or no turnout. If one adds an assumption of the sort detailed above that people have a consumption value for voting,

this would likely improve the predictive ability of the model, but it fails this condition. The assumption that voting has consumption value and can be represented by the addition of a constant to the turnout calculus is indistinguishable from a universe of other explanations for this constant (e.g., civic virtue). Further, it violates the basic spirit of model, which explicitly focused on investment and not consumption.

III. The assignment of a particular value to an assumption is not, by itself, the dispositive factor in achieving a result. One way to think of this condition is to examine whether or not there is "result convergence" (a continuity condition) when the parameter embodying the assumption is subject to perturbations. As the assumption moves closer and closer to being satisfied, the outcome of the model should move closer and closer to the final result. To the extent there are discontinuities or knife-edge results based upon changing the value of an assumption, one would question the use of the assumption, as the assumption is doing all the work, not the core logic of the model. One example of an assumption that fails to satisfy this condition is the neighborhood metric relied upon in the social science version of the Ising model (see the earlier discussion and Figure 1.1) – slightly different metrics yield dramatically different results. Assumptions are always abstractions from reality, and as such, one would want to believe that results hold within a reasonable neighborhood for each parameter value. If results only hold for a particular value and no other, one would have to defend the choice of this value or else discard the model.[16]

[16] One might fault these criteria because they cut against the spirit of Friedman's instrumentalism – if a theory succeeds in out-of-sample tests, why bother with any consideration of the assumptions? The problem, as Nagel (1961, 1963) points out in his discussion of Craig's theorem, is that theoretical statements (whether true or false) have great value in organizing knowledge and in producing new theories. All theories are not equally useful in pursuing these ends, even if they have some empirical success. Moreover, Friedman does not explicitly deal with the problem that theories must include statements (which are themselves theories) on how to map theoretical objects to empirically observable objects. To the extent that this process allows for sleight of hand, insofar as arbitrary or shifting domain restrictions are often embedded in such practices, one should place less confidence in theories that violate the constraints raised here. As a concrete example, imagine N researchers provide N competing theories, all of which are consistent with a data set A, but all of which are wrong. If a new data set B is introduced, by chance alone some subset of the N

Readers will note that one class of assumptions would deserve particular scrutiny under the above conditions: technical assumptions. The name itself should raise a note of caution! To the extent that one finds technical assumptions in a model in which slight perturbations of these assumptions drive the behavior of the model, it is difficult to separate this problem from that of spurious correlation in the world of statistical methodology. At root, when technical assumptions drive results, one has to question how comfortable one is with the idea that an assumption that is unrelated to the phenomenon in question ends up accounting for a model's brittle results. To the extent it is difficult to justify one's choice of such assumptions (or values for them) endogenous to the problem under consideration, this seems exactly the same kind of practice that leads careless empirical researchers to include independent variables in a willy-nilly fashion until an arbitrarily high R^2 is reached.

At this point, many formal theorists may object to the above conditions as overly limiting. Providing a statement along with every deductive model of how changes in the values chosen for the assumptions would impact the results of a model would be quite difficult. Game theory, for example, has no ready-to-hand theory of equivalence classes of games, and it is typically the case that any change in an assumption or parameter value results in an incommensurable game.[17] To demonstrate that results are constant across perturbations in the assumption space would thus be impossible unless one adopts a different approach to modeling.

theories may still be seen as valid. There also will be new researchers who produce new and (let us assume) wrong models that comport with (A × B). If there are continual novel datasets, one would hope that at the end of the day, all the prior models are rejected and the difficulty of inventing a new model that comports with (A × B × . . .) is progressively more difficult. Whenever data are sparse, however, it seems something else is needed to prevent random chance from usurping good judgment as the final arbiter between competing theories. Keep in mind that journals typically print only positive results, and thus condition III (equivalence classes in parameter space) is particularly useful – else, one cannot know when a model has failed over and again only to be resuscitated at the last moment by a fortuitous selection of parameter values or domain restrictions. In addition to Nagel, Simon (1963), Samuelson (1963), Boland (1979), and Hirsch and de Marchi (1984) represent high points in the lengthy debate over Friedman.

[17] Some of the inspiration for this concept comes from efforts to examine complex models for robustness. In particular, Miller (1998) was particularly influential, as well as the idea of parametric continuity from the optimization literature – see Sundaram (1996) for an overview.

But this is exactly the point. The parameter space generated by assumptions in deductive models should be viewed in exactly the same manner as the parameter spaces underlying statistical models. These "assumption spaces," as noted earlier, are usually quite large, even for simple games. To the extent we are critical of empirical modelers for loading the dice in the myriad of unreported choices they make in formulating a model, so, too, should we be critical of the choices made by formal theorists, especially as it is very difficult to know how crucial a given assumption or parameter value is in generating a result. In this sense, parsimony in parameter spaces is just as valuable in a formal model as it is in an empirical model.

One example of the importance of the concept of assumption spaces is demonstrated in Ramsay and Signorino (2003). Their goal is to derive a statistical model of the divide-the-dollar game directly from the extensive form of the game. The players each have a reservation value that is unobserved. To generate a unique maximum likelihood estimator (MLE), Ramsay and Signorino assign disturbance terms to these reservation values that are IID logistic variables. Their claim is that the MLE estimator they derive depends solely upon the form of the game; further, if one does statistical work using divide-the-dollar games as the data generating process (e.g., through experiments with human subjects), *only* their MLE estimator is appropriate. The problem they point to is important – using an Ordinary Least Squares (OLS) or some other estimator may not be appropriate to the game generating the data. But the fact that their estimator achieves different results than other statistical models should come as no surprise.

The main worry is that their assumption that the disturbance terms for each player's reservation value are IID and logistic violates condition III. For different distributions of the disturbance terms, Ramsay and Signorino would have to derive a unique MLE estimator for the divide-the-dollar game, and there is no logical implication from the structure of the game that one particular distribution is appropriate. Results generated with one assumption concerning the disturbance term would not be the same as results generated with other disturbance terms, even those with similar properties (e.g., a truncated normal bounded by 0 and 1). The fact that Ramsay and Signorino get different results with their method does not of itself cast doubt on prior work that makes different distributional assumptions. To demonstrate their

claim that their method is better, Ramsay and Signorino would have to show that the only logical choice of disturbance term was IID and logistic. Given that the subject of study is the divide-the-dollar game across different cultures, this would seem to be a difficult task that seems unrelated to the main problem under consideration (thereby violating condition II as well).[18]

Amendment 2: Logical Implications[19]

Reading this chapter might incline one to the belief that large N studies are the only appropriate way to test models. The main problem with mathematical modeling in the social sciences emphasized throughout is the disconnect between models and empirical tests that have the power to discriminate between competing models.

The reason models need to be clear about their empirical referent (i.e., the dependent variable that will test the model) is that, all too often, we resort to games such as the iterated prisoners' dilemma (IPD) and make broad claims about the results. The IPD purports to study cooperation, and surely it does detail cooperation of a kind. The most celebrated "result" of the IPD demonstrated by Axelrod's (1984) path-breaking tournaments is that tit-for-tat is the right strategy to employ when confronted by an IPD – many articles have taken this as a starting point and the literature on the IPD is vast. Unfortunately, this result is wrong on technical grounds (Binmore 1997), as the success of tit-for-tat depends upon the starting population. Another concern is more fundamental. Axelrod and other scholars use their results from the IPD to arrive at policy recommendations for phenomena ranging from Cold War deterrence strategies to regulatory compliance on the part of firms. Although results from the IPD might help one's intuition in

[18] One also should critique the assumption that the errors are IID. In many cultures, it is possible to imagine that the variance of the error term on the reservation value is correlated with the initial offer by player 1. For example, if player 1 makes a high offer (more than 50 cents, which some might label "irrational"), there would likely be very little variance in the error term. And if player 1 offers something close to 0 one would expect little variance. Across cultures, the variety of disturbance terms that might account for the data would be quite large. The logic of condition III would argue that how one's results change based on this choice should be the focus of study rather than a minor technical aside.

[19] This section was derived from conversations with Robert Keohane, during a seminar we jointly taught at Duke University on qualitative research methods in the fall of 2003.

facing the complexities of real-world problems, intuition is difficult to falsify. Once one decides that the IPD and tit-for-tat embody the essence of cooperation, everything looks like an IPD and it is difficult to know when one is in new territory.

Friedman's admonition to do out-of-sample testing curbs this kind of loose analogizing. Large N studies allow for the necessary and *repeated* confrontations with novel data that all modeling requires. There is, however, another approach to model testing that has a long history in the natural sciences as well as the social sciences (for an overview, see King, Keohane, and Verba 1994, section 1.1.3). One can derive logical implications of a model, and see if the implications in fact hold. The example provided by King, Keohane, and Verba concerns the study of dinosaur extinction:

> Nevertheless, dinosaur extinction can be studied scientifically: alternative hypotheses can be developed and tested with respect to their observable implications. One hypothesis to account for dinosaur extinction, developed by Luis Alvarez and collaborators at Berkeley in the late 1970s (W. Alvarez 1990), posits a cosmic collision: a meteorite crashed into the earth at about 72,000 kilometers an hour, creating a blast greater than that from a full-scale nuclear war. If this hypothesis is correct, it would have the observable implication that iridium (an element common in meteorites but rare on earth) should be found in the particular layer of the earth's crust that corresponds to sediment laid down sixty-five million years ago; indeed, the discovery of iridium at predicted layers in the earth has been taken as partial confirming evidence for the theory. Although this is an unambiguously unique event, there are many other observable implications. For one example, it should be possible to find the meteorite's crater somewhere on Earth. (King, Keohane, and Verba 1994, 11)

Note that unlike many modeling exercises, the researchers studying extinction did not attempt to fit a model to existing facts.[20] Failure to succeed at this activity is a sign of mathematical ineptitude, rather than a signal of a model's strength. For logical implications to be used as a test, they must be novel and uniquely connected to the logic of the model. This is a harder set of conditions then one might expect; as King, Keohane, and Verba note in a footnote on the same page:

> However, an alternative hypothesis, that extinction was caused by volcanic eruptions, is also consistent with the presence of iridium, and seems more

[20] For example, if we know that parties converge in a two-party system, producing a model that has this as an implication is trivial.

consistent than the meteorite hypothesis with the finding that all the species extinctions did not occur simultaneously.

We thus have a conundrum. It must be the case that logical implications are able to test theories – the examples of such tests are abundant and important in the history of science. It seems impossible to imagine that anyone would not have been jubilant when Sir Arthur Eddington's expedition verified Einstein's claim that mass curves space, by observing how light bends around stars. When there are other models that (as in the case of iridium deposits) might produce the same logical implication, one has to adopt a betting mentality. The test provided by deriving logical implications matters more when the implications are novel and untested – after one finds, for example, that light bends around stars, future models that "predict" this fact do not gain the same amount of credibility.

Novel implications are thus important in testing models, but one must be careful that the implication follows uniquely from the model in question. Unlike out-of-sample statistical work with large numbers of observations, logical implications, as a kind of test, are of a more qualitative nature. One has to ask how surprising the implication is, and how likely it is that a large class of other models might yield the same result. As noted in King, Keohane, and Verba, one aspect of testing a model via logical implications is shared with statistical work. To the degree possible, one should maximize variance in one's implications thereby enhancing the odds that one might be wrong. For example, instead of merely predicting the presence of iridium, one might derive a specific amount or pattern of sediment around the impact crater, decreasing the odds that other models produce the same implication.

Fortunately, this practice has a long history in natural science and we can stick with Einstein to provide a final example. In a close analogue to the Eddington experiment, the astronomer Sergei Kopeikin measured the displacement of light from a quasar moving around the mass of Jupiter during an eclipse. Because displacement depends upon gravity, he tested whether Einstein's theory that the speed of gravity is equal to the speed of light is true or not (it is – see Whitfield 2003). This test, obviously, is much easier to falsify, as any deviation from the constant for the speed of light would signal disconfirmation. An example of how to derive and test logical implications in the social sciences will be presented in Chapter 5 of this book.

LOOKING AHEAD

One of the core arguments of this text will be that deductive models are most useful in generating intuition about a problem, especially when one investigates limiting cases. But just as important is the process of developing an equivalence class where one has some idea about how changes in assumptions or parameter values change the results of a model.

Computational models supplement game theoretic models by allowing the researcher to investigate explicitly the properties of the assumption space. Computational modeling thus extends purely deductive work by generating equivalence classes of models, thereby increasing the confidence we have in our results. Of course, I agree wholeheartedly with Friedman's original statement of epistemology: Theory without an empirical referent is almost always navel-gazing. To the extent we can all agree on what the salient dependent variables are, thereby avoiding vague mappings from models to empirical referents, so much the better.

The goal of this book will be to provide the tools necessary to develop links between the three methodological traditions in the social sciences and avoid the problems detailed above. As there are already excellent texts on statistical methodology and game theory, much of the focus will be on computational models and out-of-sample forecasting. If the use of out-of-sample empirical tests are the best way to sort between competing theories, computational models naturally lend themselves to drawing out the implications of a purely deductive theory by allowing the researcher to build models with more verisimilitude, thereby decreasing the "gap" between the analytic model and the empirical test.