The Political Methodologist

Newsletter of the Political Methodology Section
American Political Science Association
Volume 4, Number 1, Spring 1991

Editor: Charles H. Franklin, Washington University

Associate Editor: Larry Bartels, University of Rochester

Contents

Notes From the Editor

Charles H. Franklin
Washington University

Every social scientist must be concerned with epistemology. While too much concern for the topic can lead to profound depression, a healthy concern serves to highlight the problem of inference in the social sciences. This issue of *TPM* is largely devoted to this subject.

The most powerful idea in science is the experiment. The experiment embodies two critically important notions: the falsifiability of hypotheses and control over both experimental and extraneous influences. Yet experiments are not always as useful as they might be. In this issue of *TPM*, we devote considerable attention to this topic. While experiments can be very valuable, as pointed out by Kinder and Palfrey, and have played a critical role in examining the nature of the survey response, as Feldman shows, there is still room for some constructive criticism of the way we handle experimental data.

In my contribution to the issue, "Efficient Estimation in Experiments", I point out that exclusive reliance on the experimental design is not an optimal approach to inference. Many, perhaps most, experiments are analyzed through analysis of variance techniques which do not encourage one to think of the complete model of behavior underlying the experiment. This leads to two unfortunate consequences. First, by omitting non-experimental influences from the model, the analyst pays a heavy price in standard errors that are larger than they need to be. As this is the focus of the article, I'll not pursue this point here.

The more troubling problem with the analysis of experiments is that the results seem to often lend themselves to post hoc rationalizing (or "theorizing", if you prefer.) I recently received a very interesting paper reporting experimental results. The design was a $2 \times 2 \times 2 \times 2$ factorial design. When a saturated ANOVA model for this experiment was specified, there were a total of 15 effects to estimate, since not only the four main effects, but also the many interaction terms were included in the analysis. Of these 15 possible effects, only one was significant at the .05 level, with four more significant at the .10 level. The unsettling thing about this paper was that the authors served up an explanation for the significant three way interaction, though no a priori arguments were made which would lead one to expect that this particular interaction would be significant. While the post hoc explanation was plausible, and could certainly be justified as a suggestion for further research, I find this very typical approach extremely troubling.

The reason I am bothered is that experiments have the great virtue that they are *designed.* They are made on purpose, for a purpose. Where a researcher may ransack a data archive looking for something interesting, an experiment by its very nature must have been designed for a specific hypothesis. Yet when analysis consists of looking at all interaction effects, and then inventing post hoc explanations, I think this negates the very power of the experimental method. Experimental design ought to force us to be much more explicit about our priors, yet the common analysis practices severely undercut this strength.

Because experiments demand that we specify the design in advance, they would seem to also demand that we have explicit predictions of the effects we expect. If the theory predicts, a priori, that there should be a third order interaction effect, then the experiment is a powerful method of testing for this. Yet I get the distinct impression that many of these high order interactions are included in the analysis with no prior expectations whatsoever. When a significant second or third order interaction effect emerges, we rush to offer intriguing "theoretical" explanations. Little consideration is given to the likelihood that what is being explained is merely the outcome of a one in twenty chance that a null hypothesis will be rejected by mistake. In doing this, I think we forsake our strongest claims to science and revert to shamanism.

It is paradoxical that the most theoretically driven design can produce the most post hoc explanations. Instead of this common practice, experimentalists should be explicit about their theoretical priors. This would strengthen both the arguments and the inferences which are drawn. By more fully specifying the theoretical model, experimentalists can make more persuasive cases and can improve the efficiency of their estimates in the process.

This issue of *TPM* is our largest ever. As editor, I am pleased to see that members of the section have chosen to submit substantive articles for publication here. I see the mission of *TPM* as one of stimulating work in the broad field of methodology, rather than simply redistributing announcements, though we do that too. Suggestions for topics of future issues are welcome, as are articles. Submission instructions appear on the back cover. —*CHF*

An Experimental Political Science? Yes, an Experimental Political Science

Donald R. Kinder
University of Michigan

Thomas R. Palfrey
California Institute of Technology

Introduction

In 1924, the question occurred to Harold Gosnell whether turnout on election day could be enhanced by encouraging citizens to vote. The question was commonplace, but what Gosnell did about it, particularly for his time, was not: he undertook an experiment (Gosnell 1927). In the summer preceding the otherwise forgettable presidential contest between Calvin Coolidge and John W. Davis, Gosnell assigned neighborhoods lying within 12 typical districts in the city of Chicago to one of two conditions (or "treatments"). Residents living in neighborhoods designated as experimental were sent postcards that pointed out voter registration deadlines and locations, and went on to suggest that citizens of Chicago who failed to exercise their sacred right to vote were little different from those "slackers" who refused to defend their country in time of war. Meanwhile, residents of utterly comparable neighborhoods assigned to the control condition were left alone. On election day, about 8 percent more of the experimental group than the control group actually turned out to vote. The answer to Gosnell's question was *yes.*

In the more than sixty years that have passed, relatively few political scientists have followed Gosnell's excellent example. Most of what political science does in the name of science has nothing to do with experimentation. Too often experiments are regarded as exotic or silly or simply irrelevant; they are what chemists do or, closer to home, what psychologists or wayward economists do, but not what we political scientists do. The science of politics, so runs the standard argument, cannot be an experimental one.

We disagree. Experiments will never dominate the study of politics; nor should they. But while an exclusively experimental political science is neither realistic nor desirable, a political science based on a variety of empirical methods, experimentation prominent among them, is both within our reach and well worth reaching for, a point we hope to suggest here.[1]

[1] This is a much abbreviated version of an essay that will appear as the introduction to Donald R. Kinder and Thomas R. Palfrey, *Experimental Foundations of Political Inquiry*, University of Michigan Press, forthcoming.

Experimentation Defined

Experimentation may refer to a single form of scientific inquiry, but as a practical matter, experiments are amazingly diverse. Experiments are undertaken in the laboratory and in the field. Experimental investigations focus on individuals, groups, neighborhoods, schools, organizations, cities. Some experiments are revelatory in aim, as in Milgram's (1974) famous demonstrations of obedience to authority. Others are carried out essentially for methodological and measurement purposes, a tradition inaugurated by Rice's (1929) experimental investigation of interviewer effects and sustained in the 1980's by a proliferation of experiments devoted to understanding the effects on the expression of public opinion due to question wording, format, and placement (see, for example, Schuman and Presser 1981; Tourangeau and Rasinski 1988). Still others are undertaken in the interest of learning about the effects of social policies, thereby heeding Donald Campbell's (1969a) plea for an "experimenting society." And in their most celebrated incarnation, experiments provide the wherewithal for testing, refining, and even, on occasion, for contradicting established theory. All of this diversity is wondrous and in many respects admirable, but it does raise the question of what is it, precisely, that we are promoting. What *is* an experiment, anyway?

For all their diversity, experiments have in common a spirit of intervention. Experiments *intrude* upon nature, and they do so (almost always) to provide answers to causal questions. This is the distinguishing feature of Gosnell's experiment, conducted so many years ago. Gosnell had a causal proposition in mind – providing prospective voters with information about registration procedures will enhance the likelihood that they will make it to the polls come election day – and tested it by intervening in the natural ongoing political process. When we carry out experiments, we are not "merely taking what comes"; rather, we are "making observations in circumstances so arranged or interpreted that we have justification for analyzing out the factors relevant to our particular inquiry" (Kaplan 1964, p. 162).

It is the feature of intervention, and the control that such intervention brings, that distinguishes experimental research from other systematic empirical methods. In the fully realized experiment, the investigator controls the *production* of settings, the *creation* of treatments, and the *scheduling* of observations. The investigator does so in order to eliminate (or at least reduce) threats to valid inference. Settings are produced in order to exclude various nuisance factors that might otherwise interfere with the causal relation of interest. Treatments are created in order to isolate precisely the causal factor (or factors) of interest. Observations are scheduled in order to reduce the likelihood that the measured effects are contaminated by other causes. In these various ways, the experimentalist intervenes in order to eliminate alternative rival interpretations, with the hope, not always realized, of being left with only a single, plausible interpretation.

In pursuit of interpretable comparisons, experiments characteristically feature both control groups (or multiple treatments) and random assignment. Multiple treatments may be created in an effort to minimize the effect of extraneous factors, or to decompose a complex phenomena, or to test theoretically-derived parametric predictions, or to explore interaction effects between key variables. In all these cases, the purpose and advantage of the experiment is to create precise and telling comparisons. Moreover, by randomly assigning subjects to treatments, the experimentalist, in one elegant stroke, can be confident that any observed differences must be due to differences in the treatments themselves (within the limitations established by statistical analysis). By sweeping aside a host of alternative interpretations, random assignment is "the great *ceteris paribus*" of causal inference" (Cook and Campbell 1979 p. 5).

Experimental Strengths

Testing Causal Propositions

No doubt experimentation's most emphasized advantage is its capacity to test cause-effect relations, a virtue that is vividly on display in Cook and Campbell's (1979) formulation of the idea of experiment itself:

> The word *experiment* denotes a test, as when one experiments with getting up two hours earlier to see if this makes one's working day more productive. The test is usually of a causal proposition: for example, does garlic or curry add a better flavor to certain rice dishes? There are some uses of the concept of experiment where the link with cause is not immediately obvious, yet still paramount. For instance, an airplane is "experimental" only if one wants to test whether it flies faster, more efficiently, or more safely than some alternative.
>
> The notion of a "trial" or deliberate manipulation is also linked to experimenting. Actually getting up earlier on some mornings is the most direct way of evaluating how one's productivity changes; using curry on some occasions and garlic at others will enable one to evaluate which seasoning improves the rice dish; and without flying the experimental airplane, it will be difficult to test (pp. 2-3).

The unrivaled capacity of experiments to provide decisive tests of causal propositions follows immediately from two aspects of control emphasized in experimental practice: the creation of treatments of interest, and the assigning of subjects to treatment conditions randomly.

One terrific example of an experimental study supplying a crisp answer to a causal question comes from research on the long standing question of how – and how well – ordinary citizens come to their views on public life. In *The Changing American Voter*, Nie, Verba, and Petrocik (1979) challenged Converse's (1964) original and powerful thesis that Americans were innocent of ideology, by demonstrating that, beginning in 1964, public opinion suddenly became more coherent and better organized. Nie, Verba, and Petrocik interpreted their results to mark a sea change in public thinking, one set in motion by the ideologically tempestuous campaign of 1964.

Perhaps. But also in 1964, the national election study tinkered with the survey questions that have played a central role in the debate over ideology. The alterations seem minor enough, but because they were introduced at the precise moment of the apparent dramatic change in American public opinion, and because we know from other research, most of it experimentally based, that ostensibly minor changes in question wording can sometimes produce sizable differences in opinion (e.g., Schuman and Presser 1981), such tinkering constitutes a rival explanation of some plausibility. In Campbell and Stanley's (1963) terminology, the observed change in public opinion between 1960 and 1964 might be due to change in *instrumentation*, nothing more.

Sullivan, Piereson, and Marcus's (1978) solution to this puzzle was to carry out an experiment. By random determination, one half of the respondents to a Twin Cities survey were asked for their views on political issues using the pre-1964 question format; the other half were questioned using the format introduced in 1964. In this elegant experimental design, the ideological character of American electoral politics is obviously held constant – all respondents are interviewed at the same time in the same political setting – while only the opinion assessment technique is systematically manipulated.

It turns out that the experimental manipulation of question format produced large differences in the pattern of relationships between opinions on government policy, differences that mimic in a remarkably fine-grained way the differences that Nie and his associates had reasonably attributed to transformations in the nature of American politics. In light of these experimental results, most if not all the change in the structure of public opinion observed in 1964 now appears to be artificial, induced not by alterations in politics but by mundane modifications in question wording: a pure and horrifying example of instrumentation change masquerading as real change.

Analytic Decomposition

By creating treatment and control conditions, the experimentalist is able to isolate a single causal variable at a time. Put another way, experimentalists need not wait for natural processes to provide crucial tests and telling comparisons: they can create them on their own. Consider, for example, Issac, Walker and Thomas's (1984) experimental investigation of the obstacles that stand in the way of the decentralized provision of collective goods. The greatest difficulty, of course, is the inclination among individuals to "free ride": to benefit from the collective good without bearing the cost of its provision. The experiment reported by Isaac and his colleagues focused on two factors that might affect the extent to which free riding occurs: the private gain to be realized from free riding, and the size of the group. In naturally occurring groups, these two factors are difficult to disentangle and the first is hard to measure. Enter experimentation. Isaac, Walker, and Thomas show that while private incentives operate in the expected direction, the effect of group size by itself is just the reverse of the conventional wisdom: as group size increases, free riding *declines*.

Interdisciplinary Ties

Exchange across disciplinary boundaries is almost always slow and uncertain. By providing a common if not quite universal scientific language, experimentation promises to speed this process up. Especially likely are interdisciplinary collaborations between political scientists, on the one hand, and economists and psychologists, on the other – as in the burst of collaborative experimental research on collective action (Dawes 1980).

In the experimental approach, investigators often go to great lengths to identify and isolate the effects of specific variables, holding everything else as constant as possible. Consequently it is quite natural that the theoretical models used in conjunction with experimental data are ones that systematically and rigorously formalize the effects of one set of specific variables on others. Furthermore, theories that are highly parametric and formalized generate predictions that are often too detailed and precise to test appropriately with data other than experimental, either because certain key variables are unmeasurable, or because there are simply too many conditions varying simultaneously. This produces a natural synergy between formal, positive theories of politics, imported largely from economics, and a brand of experimental investigation that also has its origins in economics. New ways to study political processes from both the theoretical and the empirical side are the happy result (e.g., McKelvey and Ordeshook 1986; Boylan *et al.* 1991).

In a similar way, the importation of concepts and experimental techniques from psychology, especially to study individual information processing and decision making, is having a significant impact on how we assess the capacity of citizens to live up to the burdens of democracy (Lodge, McGraw, and Stroh 1989; Quattrone and Tversky 1988). Indeed, the very meaning of opinion is currently being revised, provoked by new findings and theories in cognitive psychology (Tourangeau and Rasinski 1988; Zaller and Feldman 1990; Kinder and Sanders 1990). An under-appreciated but most attractive quality of an experimental political science is its interdisciplinary appetite.

"Stubborn Facts" and Theoretical Invention

Empirical results are often the parent of theoretical invention. When results challenge orthodox understandings, and when they cannot be dismissed, they may lead to real advances. Like Cook and Campbell, we:

> ...find much to value in the laboratory scientist's belief in "stubborn facts" that "speak for themselves" and which have a firm dependability greater than the fluctuating theories with which one tries to explain them. Modern theorists of science – Popper, Hanson, Polanyi, Kuhn, and Feyerabend included – have exaggerated the role of comprehensive theory in scientific advance and have made experimental evidence almost irrelevant. Instead, exploratory experimentation unguided by formal theory, and unexpected experimental discoveries tangential to whatever theory motivated the research, have repeatedly been the source of great scientific advances, providing the stubborn, dependable, replicable puzzles that have justified theoretical efforts at solution (1979, p. 24).

It is no accident that Cook and Campbell refer explicitly to *experimental* discoveries. When experimental discoveries produce anomalies, they are less apt to be dismissed. Why? It is impossible to replay history and expensive to redo public opinion surveys. But experiments can be replicated, and by custom, they are. The experimental discovery of anomalous results, which survives repeated replications, is more likely to be taken seriously, and may lead in time to better theory.

A conspicuous example of replicated anomalies of high relevance to political science can be found in the experimental work of Kahneman and Tversky. In a series of ingenious experiments, Kahneman and Tversky have uncovered a catalogue of systematic departures from "rational" decision making under uncertainty (e.g., Tversky and Kahneman 1981). Such discoveries, which are anomalous from the perspective of orthodox rational choice theory, have had a sensational impact on theory and research on decision making throughout the social sciences (for a partial review, consult Abelson and Levi 1985). Because the basic results have proven robust (Grether and Plott 1979; Loomes, Starmer, and Sugden 1991), a good bit of theoretical invention has followed (e.g., Thaler 1980; Machina 1982; Kahneman and Tversky 1979). The modification of utility theory to accommodate experimental observations is currently one of the most active areas of research in all of mathematical economics.

Flexibility Across Levels of Aggregation

In principle and in practice, the experimental method applies widely across different levels of aggregation. Few would dispute that it is desirable for an empirical method to have a firm footing at the individual level, yet lend itself to empirical analysis across a broad spectrum. This is especially so in political science, where relevant applications exist at all levels of aggregation, from the study of individual citizens trying to make sense of campaigns to the analysis of bilateral negotiations between two superpowers. Experiments have the flexibility to test theories and to provide empirical insights at all levels.

Consider, as illustrations of the experimental range, the investigation of individual judgment and choice (e.g., Iyengar and Kinder 1987); the examination of institutional rules for committee and legislative behavior (e.g., Fiorina and Plott 1978; Eavey and Miller 1984); experiments on collective action, some of which are explicitly concerned with studying how behavior changes as groups get larger (e.g., Isaac and Walker 1989); and the experimental probing of the interplay between candidates and voters through the democratically critical technique of elections (e.g., McKelvey and Ordeshook 1987).

Experimental Shortcomings

Experiments, like other methods, have liabilities as well as strengths: the ambiguity surrounding the meaning of experimental treatments; the practical difficulties in applying experimental methods to some problems in political science; and the hazards that can threaten experimental generalization. Of the three, we confine our attention here to the problem of generalization, both because it occupies such a prominent place in arguments against experimentation in political science, and because such arguments seldom rise much above sneering references to college sophomores, on whose backs many experiments run.

To generalize from particular experimental arrangements and populations to the real political world is to participate in what Campbell (1969b) has called, for dramatic effect, "the scandal of induction." Always and inescapably, generalizations are matters of opinion. Concern about the generalizability of experimental results in particular usually takes one of three forms. First, because experimental participants ordinarily know that they are taking part in the study of something (even if they're not sure what), this knowledge alone may induce alterations in their behavior. Second, experiments are often conducted with samples of convenience, leading to skepticism over whether experimental results can be generalized safely to the populations of real interest. For American social scientists situated in universities, no population is of course more convenient than the local student body. And the typical college sophomore, as Hovland (1959) warned some years ago, and as Sears (1986) has recently documented, may be a rather peculiar creature. Third, experimental results are always subject to the charge that they depend precariously on exactly how the independent variables were created. These concerns about generalizing across settings, populations, or treatments almost inevitably accompany the presentation of experimental results, and so they should.

The most effective remedy for the problem of experimental generalization is to carry out *selective* replications (Carlsmith, Ellsworth, and Aronson 1976). We emphasize selective because we have no interest, and the field has nothing like the required resources, to follow a program of *comprehensive* replication. By no means are we suggesting that to be assured about the generalizability of a particular experimental result, we must have in hand corroborating evidence from perfectly representative samples of populations, settings, and treatments. Our advice instead is to pursue carefully chosen and selective replications. The point is to vary settings, populations, and treatments in ways that represent revealing and usually difficult tests of generalization. In this way, we *probe* the generalizability or robustness of experimental results.

For reasons of control and convenience, most experiments will no doubt continue to be undertaken in artificial settings, with college student subjects confronting treatments that have no exact counterpart outside the laboratory. But these can be usefully complemented by occasional experimental ventures that place a higher premium on matters of external validity. Happily, there appears to be a fair number of such experiments already completed in political science. For example, Cover and Blumberg (1982) investigated the celebrated incumbency advantage enjoyed by members of the U.S. House of Representatives by withholding the flow of congressional mail from some constituents and not from others. Or consider Levine and Plott's (1977) field experiment of agenda influence, or Fiorina and Plott's instructive result (1978) that the process of committee decision making takes a very different path when real, material incentives are at stake, or Weiss's (1982) experiment on complexity and decision making among flesh and blood policy analysts. Taken together, these various examples testify that the generalization problem is no reason to give up on experiments, that the generalizability of results can be systematically probed within the experimental method itself.

Moreover, in one class of experiments, the question of the generalizability or the robustness of results, while interesting, is not critical to the value of the experiment. We think of these as "demonstration" experiments. Many of the Kahneman and Tversky studies exposing breaches of utility theory fall into this category. Such experiments demonstrate the *existence* of choices where individuals consistently (and replicably) violate some of the axioms on which utility theory is based. They pointedly do not claim that individuals *always* violate these axioms. In fact, much of their work entails specifying those environments where such violations habitually occur. Another example comes from recent experiments in economics, which demonstrate, against the predictions of theory, the *possibility* of bubbles and crashes in stock markets (Smith, Suchanek, and Williams 1988). Again, the issue of robustness is of some interest here, but the main point is to challenge theory through a display of unambiguously anomalous evidence.

In sum, while the risks in generalizing from experimental results may never be eliminated entirely, they can be sharply reduced: by diminishing, or bypassing altogether, the artificiality of the experimental setting through field experiments; by extending experimental tests to diverse or difficult samples; by creating treatments that are representative of real settings; and more. Through ingenuity, opportunism, and sheer effort, the "scandal of induction" becomes just another puzzle, no different in kind from familiar problems of design, measurement, and analysis.

More Experiments!

Here and there across the discipline of political science we detect welcome signs of a growing interest in and sophistication about experimentation. And so there should be. The advantages that experimentation provides – testing causal propositions, unpacking complexity, accelerating interdisciplinary conversations, turning up replicable "stubborn facts", moving smoothly across different levels of aggregation – should make experimentation practically irresistible. And if the scientific justification for an experimental political science is unpersuasive, consider this: we have found that actually doing experiments – planning them, carrying them out, analyzing them, puzzling over the results – is often great fun.

References

Abelson, Robert P., and Ariel Levi (1985). Decision Making and Decision Theory. In Gardner Lindzey and Elliot Aronson, eds., *Handbook of Social Psychology.* New York: Random House.

Boylan, Richard, John Ledyard, Arthur Lupia, Richard D. McKelvey, and Peter Ordeshook (1991). Political Competition in a Model of Economic Growth: An Experimental Study. In Thomas R. Palfrey ed., *Laboratory Research in Political Economy.* Ann Arbor, Michigan: University of Michigan Press.

Campbell, Donald T. (1969a). Reforms as Experiments. *American Psychologist* 24:409-429.

Campbell, Donald T. (1969b). Prospective: Artifact and Control. In Robert Rosenthal and Robert Rosnow, eds., *Artifact in Behavioral Research.* New York: Academic Press.

Campbell, Donald T., and Julian C. Stanley (1966). *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Carlsmith, J. Merrill, Phoebe C. Ellsworth, and Elliot Aronson (1976). *Methods of Research in Social Psychology.* Reading, Massachusetts: Addison-Wesley.

Converse, Philip E. (1964). The Nature of Belief Systems in Mass Publics. In David E. Apter, ed., *Ideology and Discontent.* New York: The Free Press.

Cook, Thomas D., and Donald T. Campbell (1979). *Quasi-Experimentation.* Chicago: Rand McNally.

Cover, Albert D., and Bruce S. Brumberg (1982). Baby Books and Ballots: The Impact of Congressional Mail on Constituent Opinion. *American Political Science Review* 76:347-359.

Dawes, Robyn M. (1980). Social Dilemmas. *Annual Review of Psychology* 31:169-93.

Dawes, Robyn M., John M. Orbell, Randy T. Simmons, and Alphons J.C. van de Kragt (1986). Organizing Groups for Collective Action. *American Political Science Review.*

Eavey, Cheryl L., and Gary J. Miller (1984). Bureaucratic Agenda Control: Imposition or Bargaining? *American Political Science Review* 78:719-33.

Ferejohn, John, Robert Forsythe, Roger Noll, and Thomas R. Palfrey (1982). An Experimental Examination of Auction Mechanisms for Discrete Public Goods. In Vernon L. Smith, ed., *Research in Experimental Economics, Volume 2.* Greenwich, Connecticut: JAI Press. 2:175-99.

Fiorina, Morris P., and Charles R. Plott (1978). Committee Decisions Under Majority Rule. *American Political Science Review* 72:575-98.

Gosnell, Harold F. (1927). *Getting Out the Vote: An Experiment in the Stimulation of Voting.* Chicago: University of Chicago Press.

Grether, David M., and Charles Plott (1979). Economic Theory of Choice and the Preference Reversal Phenomenon. *American Economic Review* 69:623-638.

Hovland, Carl I. (1959). Reconciling Conflicting Results Derived from Experimental and Survey Studies of Attitude Change. *American Psychologist* 14:8-17.

Isaac, R. Mark, James M. Walker, and Susan H. Thomas (1984). Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations. *Public Choice* 43:113-49.

Isaac, R. Mark, James M. Walker, and Arlington W. Williams (1989). Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Very Large Groups. Unpublished manuscript, Economics Department, Indiana University.

Iyengar, Shanto, and Donald R. Kinder (1987). *News that Matters.* Chicago: University of Chicago Press.

Kahneman, Daniel, and Amos Tversky (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47:263-291.

Kaplan, Abraham (1964). *The Conduct of Inquiry.* San Francisco: Chandler Publishing Company.

Kinder, Donald R., and Lynn M. Sanders (1990). Mimicking Political Debate with Survey Questions: The Case of White Opinion on Affirmative Action for Blacks. *Social Cognition* 8:73-103.

Levine, Michael E., and Charles R. Plott (1977). Agenda Influence and Its Implications. *Virginia Law Review* 63:561-604.

Loomes, Graham, Chris Starmer, and Robert Sugden (1991). Observing Violations of Transitivity by Experimental Methods. *Ecomometrica* 59:425-439.

Machina, Mark (1982). "Expected Utility" Analysis without the Independence Axiom. *Econometrica* 50:277-323.

McKelvey, Richard D., and Peter C. Ordeshook (1987). A Decade of Experimental Research on Spatial Models of Elections and Committees. California Institute of Technology, Social Science Working Paper 657.

McKelvey, Richard, and Peter Ordeshook (1990). Information and Elections: Retrospective Voting and Rational Expectations. In Kuklinski, J., and John Ferejohn, eds., *Information and Democratic Process.* Urbana-Champaign: University of Illinois Press.

Lodge, Milton, Kathleen M. McGraw, and Patrick Stroh (1989). An Impression-Driven Model of Candidate Evaluation. *American Political Science Review* 83:399-419.

Milgram, Stanley (1974). *Obedience to Authority.* New York: Harper and Row.

Nie, Norman H., Sidney Verba, and John R. Petrocik (1979). *The Changing American Voter.* Cambridge: Harvard University Press.

Quattrone, George A., and Amos Tversky (1988). Contrasting rational and psychological analyses of political choice. *American Political Science Review* 82:719-736.

Rice, S.A. (1929). Contagious Bias in the Interview: A Methodological Note. *American Journal of Sociology* 35:420-23.

Schuman, Howard, and Stanley Presser (1981). *Questions and Answers in Attitude Surveys. Experiments on Question Wording, Form and Context.* New York: Academic Press.

Sears, David O. (1986). College Sophomores in the Laboratory: Influence of a Narrow Data Base on Social Psychology's View of Human Nature. *Journal of Personality and Social Psychology* 51:515-530.

Smith, Vernon L., Gerry L. Suchanek, and Arlington W.

Williams (1988). Bubbles, Crashes, and Endogenous Expectations in Experimental Spot Asset Markets. *Econometrica* 56:1119-1151.

Sullivan, John L., James E. Piereson, and George E. Marcus (1978). Ideological Constraint in the Mass Public: A Methodological Critique and Some New Findings. *American Journal of Political Science* 22:233-249.

Thaler, R. (1980). Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior and Organization* 1:39-60.

Tourangeau, Roger, and Kenneth A. Rasinski (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin* 103:299-314.

Tversky, Amos, and Daniel Kahneman (1981). The Framing of Decisions and the Psychology of Choice. *Science* 211:453-458.

Weiss, Janet (1982). Coping with Complexity: An Experimental Study of Public Policy Decision-Making. *Journal of Policy Analysis and Management* 2:66-87.

Zaller, John, and Stanley Feldman (1990). Answering Questions vs. Revealing Preferences: A Simple Theory of the Survey Response. Unpublished manuscript, Department of Political Science, State University of New York at Stony Brook.

What Do Survey Questions Really Measure?

Stanley Feldman
SUNY–Stony Brook

> As I see it, a measurement model worthy of the name must make explicit some conceptualization—at least a rudimentary one—of what goes on when an examinee solves test problems or a respondent answers opinion questions; and it must incorporate a rigorous argument about what it *means* to measure an ability or attitude with a collection of discrete and somewhat heterogeneous items. (Duncan 1984, p. 217)

Survey data have long been a staple of behavioral research in political science, sociology, psychology and even economics. The origins of the social survey can be traced back almost a hundred years (Converse 1987). Since the beginnings of formal attitude measurement over 60 years ago, social scientists have devised survey questions to measure almost every imaginable mental entity in almost every conceivable subject area. The combination of standardized attitude questions and the large-scale probability sample enables researchers to collect information about recall of past behavior and subjective states from representative samples of almost any population. Survey data also have now come to rival census data as an integral part of politics and social life.

Since survey data have been so central to empirical research in the social sciences for so many years it would be reasonable to assume that the properties of survey questions are well understood. To the contrary, the closer you look at methodological research on survey questions the more embarrassing the state of our knowledge appears. Many of us have a schizophrenic relationship with survey data: we are aware of (at least some of) the literature on the problematic nature of survey data but we then conveniently ignore those problems when we use survey data to estimate models of public opinion and vote choice. Researchers have generally been content to use survey responses as direct indicators of unobserved mental constructs, recognizing that survey questions may be affected by random and systematic measurement error, but only occasionally attempting to deal explicitly with the consequences of these errors for parameter estimation.

The Fragility of Survey Responses

Researchers have long recognized that responses to attitude and opinion questions are subject to the influence of relatively minor changes in wording, question order, and response options (see Rugg 1941). In recent years, there has been an increasing amount of research using experimental designs (split- half surveys) to study the effects of questionnaire structure on survey responses (see for example Schuman and Presser 1981). Although many of these "response effects" result in small changes in marginal distributions, the effects can sometimes be substantial. Perhaps the best documented case comes from a pair of questions on communist and American newspaper reporters (Schuman and Presser 1981). When asked "Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?" 37% (in 1948) said yes. When this question was preceded by one asking if American reporters should be allowed to report the news from communist countries like Russia, yes responses to the communist reporters question increased to 73%. Response effects are almost never this large but they are frequently observed.

Many careful experimental studies have illustrated the consequences of question wording, question placement, types of response options and other factors on marginal and multivariate distributions. But, thus far, this literature has failed to develop general principles that can be used to predict when such effects will occur. Often, context effects seem to be question specific and sometimes even seemingly reliable effects fail to replicate from one experiment to another (Schuman and Presser 1981, pp. 317- 324). Converse and Presser (1986, p. 41) note that "Even small changes in wording can shift the answers of many respondents, but it is frequently difficult to predict in advance whether a wording change will have such an effect." There is no question that this experimental research has contributed to our understanding of survey questions and public opinion. However, after having examined "several hundred experiments in 344 surveys" Schuman and Presser (1981, p. 307) conclude: "Experiments do not always produce interpretable results, nor always cumulate to yield general conclusions, nor always even replicate well."

A second embarrassing problem with answers to survey questions is response inconsistency in panel data. Political scientists are well aware of this phenomenon as a result of Converse's (1970) analysis of the 1956-58-60 election studies panel and the literature that has responded to his conclusions. What is the meaning of the large random component that is apparently associated with most political issue questions? From several data sets we know that there is little or no systematic attitude change on most issues (but see Smith 1984). There is, however, a stochastic component that is as large as 60 percent of the observed variance of some questions (Achen 1975; Erikson 1979).

The problem is that different sets of (untested) assumptions lead to alternative statistical models that yield very different conclusions. Two interpretations are prominent in the literature: nonattitudes (attitude crystallization) and measurement error. Each interpretation generates a statistical model that appears to fit the data well (see for example Erikson 1979; Feldman 1989; Brody 1986). Yet neither model is derived from a theoretical understanding of the survey response process. The nonattitudes model simply assumes that people with real attitudes will give error free responses to survey questions. Temporal inconsistency thus can be attributed to nonattitudes. The measurement error model assumes that the stochastic component of the responses, i.e., that part of the temporal variation that deviates from a simple model of true attitude change, is random measurement error as defined in the test theory literature: normally distributed response error around the true score.

While both models appear to fit the data well, there is empirical evidence that contradicts both accounts. Contrary to the assumptions of the measurement error model, response inconsistency *is* related to measures of political information and sophistication (Norpoth and Lodge 1985; Feldman 1989; Zaller 1990). Contrary to the nonattitudes model, response inconsistency is *not* directly predicted by attitude centrality or salience (Feldman 1989). And response error is still substantial even among those highest in political information (Norpoth and Lodge 1985; Feldman 1989; Zaller 1990). Goodness-of-fit statistics do not appear to be of much assistance here. Not only do two very different models fit the same data well, but *neither* appears to survive straightforward tests of key predictions.

The nonattitudes (or attitude crystallization) model now seems to fail in another important respect: it does not successfully predict susceptibility to response effects in surveys. Despite obvious predictions that context effects in surveys ought to be much less pronounced among those with crystallized attitudes than among those likely to hold nonattitudes, two reanalyses of many experiments failed to find substantial evidence for the hypothesis (Krosnick and Schuman 1988; Bishop 1990). The measurement error model is of no help at all in understanding response effects. If this model makes any prediction at all, it is that there should be no systematic response effects since variations from true attitudes are simply random. With almost perfectly stable true scores over four year intervals, how can properties of the measurement instrument produce significant changes in responses? Even if one of these models were to successfully explain response instability we would apparently need at least one other model to account for the other sources of variability in survey responses.

Much statistical firepower has been directed at the response instability issue and hundreds of experiments have been designed to shed light on the problem of response effects. Is it likely that more work of this type will provide the understanding of responses to survey questions that has not yet emerged from all this research? The answer seems to be no. In the conclusion of their analysis of experimental work on response effects, Schuman and Presser (1981, p. 313) argue: "What is needed most is theoretically directed research, but exactly what this means is not so clear." Although Schuman and Presser were skeptical of the utility of theories in cognitive and social psychology for directing research on the nature of the survey response, an impressive body of research has developed in the 10 years since Schuman and Presser's book that draws directly on those theories. This research provides a very different understanding of the survey response than the implicit model that underlies much of the empirical work that uses survey data. In the end, it also may require that we alter our view of the nature of public opinion.

Toward a Theory of the Survey Response

The standard view of the survey response is that when a respondent is asked a question, he or she simply recalls the "true score" required by the question (if it does exist) and selects a response option consistent with that attitude. Should we believe that this is an accurate representation of the process of answering survey questions? It would mean that people, informed and uninformed alike, carry around in their heads preformed answers to almost all the questions that survey researchers can construct. Every time a survey question is asked, the respondent would have a response to the question already stored in memory. Even if this were true, we also have to believe that respondents can quickly recall all those preformed responses under the time pressure and low relevance of a typical survey. This does not seem very plausible.

If the face validity of this "model" appears suspect, recent attempts to construct cognitive models of the survey response also cast doubt on it. These models start by assuming that answering a survey question is in principle no different from answering any other type of question. Thus, responding to a question about George Bush's handling of the Gulf war is structurally no different from answering a question in Trivial Pursuit or responding to a question from a friend on whom the best pitcher in baseball is. This approach forces us to consider exactly what respondents are doing when we ask them to answer our survey questions. It has become the norm in discussions of cognitive models of the survey response to divide the response process into four steps: interpreting the question, information search, formulating the answer, and selecting the appropriate response. In practice, it is not at all clear that these are discrete steps. It is nonetheless useful to maintain these distinctions for analytic purposes. Several detailed discussions of these new models of the survey response are available (see Tourangeau 1984; Hippler, Schwarz and Sudman 1987; Tourangeau and Rasinski 1988; Zaller and Feldman 1988). It is not possible to provide anything like a comprehensive summary of the theoretical and experimental literature relevant to these models in the space available here. Instead, I will try to highlight the most significant aspects of the response process and discuss their implications for understanding answers to survey questions.

The distinctive qualities of these models of the survey response are generated by assumptions about question interpretation and information search. In short, interpreting a survey question and retrieving information to answer it are constructive processes that are inherently stochastic.

Consider the following survey question: How difficult is it to obtain drugs in this neighborhood? Embedded in a survey on crime, respondents will likely interpret it as asking about the ease of buying crack in a nearby alley. In a health survey the same question would generate responses about the location of the nearest drug store (Strack and Martin 1987). The existence of ambiguity in survey questions is probably a fact of life. Language simply cannot communicate a single unambiguous message (Graber 1976).

How do respondents deal with this?

A large body of work (see for example Anderson et al. 1970) shows how broad interpretative frameworks—often labeled frames, scripts, or schemata—strongly influence the way people understand the meaning of a text. This research shows that there are typically multiple interpretations of a text (or question) and that one, and only one, interpretation is activated to make sense of textual material. Thus, when respondents hear or read a survey question, the first thing they do (most likely automatically) is to activate an interpretative framework to make sense of the question. Respondents can draw from multiple interpretations but only one is activated at any time.

Given an interpretation of the question, how does the respondent go about answering the question? An obvious strategy is to search for the preformed judgment specified in the question. That is, directly recall the opinion or attitude required. It was previously argued—based on face validity—that this is an unrealistic model to apply to most respondents on many survey questions. A more direct treatment of mechanisms of memory and recall reinforces this conclusion.

Most models of memory distinguish between long-term and short-term (or working) memory. Long-term memory has almost unlimited capacity to store information. Only a very small portion of the contents of long-term memory can be accessed at any time; short-term memory contains that small portion of long-term memory that we are consciously aware of at any given moment. (For a good introduction to associative memory models see Hastie 1986.)

There are two key points about memory retrieval that we need to consider. First, since long-term memory is very capacious, it can be difficult to quickly locate any specific bit of information unless the connection between the cue used to initiate memory search and the information is very strong. Second, retrieval from long-term memory is probabilistic (Raaijmakers and Shiffrin 1980). For any particular cue, there is a set of probabilities that define the likelihood of retrieving information associated with that cue. These characteristics of memory retrieval lead to a conclusion that is contrary to an assumption of direct recall of responses to survey questions. As Reeder (1982: 252) argues, "fact retrieval (trying to find an assertion in memory) is often less efficient than computing plausibility (or inferring) and it is not always the first strategy employed in sentence verification." Reeder further asserts that:

> In everyday life it is unlikely that all facts or even the majority of facts on which people are queried are directly stored in memory. Further, memory is a rich, highly redundant store of information. Searching for any specific proposition may not be much easier than searching for a needle in a haystack. Therefore, it is often faster to select the first few relevant facts found in memory (and compute the answer) than to continue to search until an exact match can be found (1982, p. 252).

The nature of the typical survey interview almost certainly encourages quick responses and, therefore, fast retrieval strategies and very brief memory searches. This decreases the likelihood that long memory searches for the specific attitude or belief will take place and suggests that respondents probably will answer survey questions on the basis of the first thing or things that they recall (see Zaller and Feldman 1988).

The Nature and Measurement of Public Opinion

There are several important conclusions that emerge if the cognitive response framework is taken seriously. The most obvious implication is that survey questions are inherently noisy measures. Stochastic variation is introduced first through the availability of multiple interpretations of the question and then through an incomplete and probabilistic memory search. "Better questions" will never eliminate this variability.

This stochastic variation is not simply a problem for opinion measurement; it reshapes our basic understanding of the nature of opinions. Respondents do not answer opinion questions by directly recalling their opinion. Indeed, in the cognitive response models opinions exist not as individual mental representations but as distributions of *considerations*—the attitudes, beliefs, values and information that can be retrieved in response to a survey question. Thus, typical survey methods that estimate the central tendency of this distribution tell only part of the story. It is also necessary to estimate the shape of the distribution. There is as much or more information in the shape of the distribution of considerations as in the central tendency. Yet when we do obtain information about the shape (variance) of the distribution it is typically labeled error variance and considered nothing more than a threat to proper parameter estimation.

How does this framework address the problems of response instability and survey response effects? Response instability should derive in large part from the probabilistic nature of the response process. Answers to an opinion question will vary across administrations except under two conditions: the respondent does have a fixed opinion on an issue that can be consistently retrieved under survey conditions or all the considerations relevant to the survey question have identical implications. Instead of a dichotomous distinction between attitudes and nonattitudes, it is probably more useful to consider the variance of considerations with respect to an issue. As the considerations become more homogeneous, responses to survey questions look like Converse's conception of real attitudes. The opinions of some activists on the abortion issue may approximate this condition. At the other extreme, very heterogeneous considerations will behave like nonattitudes. It is likely that most people's opinions on most issues will fall between these two extremes.

To explain response effects an additional aspect of memory retrieval must be considered: what determines the probability that a consideration will be retrieved in response to a specific survey question? Psychologists use the concept of priming to refer to the case in which the probability of retrieval of a memory representation is systematically increased. For example, frequency and recency of use will increase the probability that a consideration will be retrieved in the future (Bodenhausen and Wyer 1987). Thus, anything that affects the probability that certain considerations will be recalled will influence the observed opinions, even without actual changes in the considerations.

Priming is an important component of a full model of the survey response because it helps to explain context effects in surveys and opinion change in the real world through a single mechanism. Changes in question wording and order may affect survey responses by priming certain interpretations or considerations, making it more likely than otherwise that they will influence the survey response. Experimental studies of response effects (Tourangeau and Rasinski 1988; Tourangeau et al. 1989) have found evidence consistent with priming effects. And Kinder and Sanders (1990) argue that the effects of changes in question wording can mimic the effects of political debate on public opinion by altering the considerations that respondents use to generate an opinion.

Context effects in surveys that typically get labeled methodological artifacts are thus similar in structure to the priming of opinions by politicians and the mass media. In an important substantive example, Iyengar and Kinder (1987) show that even small changes in news coverage can alter the likelihood that certain considerations will be retrieved when people are asked to evaluate the president. Priming alters evaluations of the president even though "beliefs" about the president may be unaffected by the media. Many context effects in surveys may work the same way.

The distribution of considerations that generates opinion responses is also important from the perspective of politics and the shaping of public opinion. It is much easier to understand how "opinions" may change when seen from the perspective of the retrieval of heterogeneous considerations than by assuming that issue preferences are single fixed values. The connection between priming in surveys and the political world suggests that response effects in surveys are opportunities to study public opinion. Once we recognize that an "opinion" on an issue is generally a range of possible reactions rather than a single point, multiple questions and question orders can be important tools in examining the entire distribution of opinion on an issue.

There is much more to be said about this approach to understanding survey responses than I have the space to address here. Much more empirical work is also necessary to determine the ability of these cognitive models to account for the properties of survey responses. Regardless of the outcome of future research, a major virtue of this approach is that it finally forces us to explicitly consider what a model of the survey response might look like and what its implications are for the measurement and understanding of public opinion.

References

Achen, Christopher H. 1975. "Mass Political Attitudes and the Survey Response." *American Political Science Review* 69: 1218-1223.

Anderson, Richard C., Reynolds, Ralph E., Schallert, Diane L. and Goetz, Ernest T. 1977. "Frameworks for Understanding Discourse." *American Educational Research Journal* 14: 367-381.

Bishop, George. 1990. "Issue Involvement and Response Effects in Public Opinion Surveys." *Public Opinion Quarterly* 54: 209-218.

Bodenhaussen, Galen V. and Robert S. Wyer. 1987. "Social Cognition and Social Reality: Information Acquisition and Use in the Laboratory and the Real World." In *Social Information Processing and Survey Methodology*, ed. Hans-J. Hippler, Norbert Schwarz and Seymour Sudman. New York: Springer-Verlag.

Brody, Charles A. 1986. "Things Are Rarely Black and White: Admitting Gray into the Converse Model of Attitude Stability." *American Journal of Sociology* 92: 657-677.

Converse, Jean M. 1987. *Survey Research in the United States*. Berkeley: University of California Press.

Converse, Jean M. and Stanley Presser. 1986. *Survey Questions*. Beverly Hills: Sage.

Converse, Philip E. 1970. "Attitudes and Nonattitudes: Continuation of a Dialogue." In *The Quantitative Analysis of Social Problems*, ed. Edward R. Tufte. Reading, MA: Addison-Welsley.

Duncan, Otis Dudley. 1984. *Notes on Social Measurement*. New York: Russell Sage Foundation.

Erikson, Robert S. 1979. "The SRC Panel Data and Mass Political Attitudes." *British Journal of Political Science* 9: 89-114.

Feldman, Stanley. 1989. "Measuring Issue Preferences: The Problem of Response Instability." *Political Analysis* 1: 25-60.

Graber, Doris A. 1976. *Verbal Behavior and Politics*. Urbana, IL: University of Illinois Press.

Hastie, Reid. 1986. "A Primer of Information-Processing Theory for the Political Scientist." In *Political Cognition*, ed. Richard R. Lau and David O. Sears. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hippler, Hans-J., Norbert Schwarz and Seymour Sudman. 1987. *Social Information Processing and Survey Methodology*. New York: Springer-Verlag.

Iyengar, Shanto and Donald R. Kinder. 1987. *News That Matters*. Chicago: University of Chicago Press.

Kinder, Donald R. and Lynn M. Sanders. 1990. "Mimicking Political Debate with Survey Questions: The Case of Opinion on Affirmative Action for Blacks." *Social Cognition* 8: 73-103.

Krosnick, Jon A. and Howard Schuman. 1988. "Attitude Intensity, Importance, and Certainty and Susceptibility to Response Effects." *Journal of Personality and Social Psychology* 54: 940-952.

Norpoth, Helmut and Milton Lodge. 1985. "The Difference Between Attitudes and Nonattitudes in the Mass Public: Just Measurement?" *American Journal of Political Science* 29: 291-307.

Raaijmakers, Jeroen G. and Richard M. Shiffrin. 1980. "SAM: A Theory of Probabilistic Search of Associative Memory." *The Psychology of Learning and Motivation* 14: 207-262.

Reeder, Lynne M. 1982. "Plausibility Judgments Versus Fact Retrieval: Alternative Strategies for Sentence Verification." *Psychological Review* 89: 250-280.

Rugg, D. 1941. "Experiments in Wording Questions: II." *Public Opinion Quarterly* 5: 91-92.

Schuman, Howard and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Smith, Tom W. 1984. "Nonattitudes: A Review and Evaluation." In *Surveying Subjective Phenomena*, Vol. 2, ed. Charles F. Turner and Elizabeth Martin. New York: Russell Sage Foundation.

Tourangeau, Roger, Kenneth Rasinski, Norman Bradburn and Roy D'Andrade. 1989. "Carryover Effects in Surveys." *Public Opinion Quarterly* 53:495-524.

Tourangeau, Roger and Kenneth Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103: 299-314.

Tourangeau, Roger. 1984. "Cognitive Science and Survey Methods." In *Cognitive Aspects of Survey Methodology*, ed. T.B. Jabine, M.L. Straf, J.M. Tanur and R. Tourangeau. Washington, D.C.: National Academy Press.

Zaller, John and Stanley Feldman. 1988. "Answering Questions vs. Revealing Preferences: A Simple Theory of the Survey Response." Paper presented at the 1988 meeting of the American Political Science Association.

Efficient Estimation in Experiments

Charles H. Franklin
Washington University

The classic experimental design, with random assignment to treatment and control, is an elegant solution to the problem of inference in the social sciences. The power of experiments comes from their ability to guarantee that no other variables are correlated with the treatment. This means that even if a host of other influences affect the response, we can still estimate the experimental effects without bias. This is a profound and beautiful discovery. Still, it is too easy to celebrate this miracle while overlooking some practicalities. In this article I want to show why experiments "work" in the first place, and how we can increase the efficiency of our estimates of experimental effects.

Why Experiments Work

Let us take as the prototypic case a simple design with N subjects who are randomly assigned to either treatment or control groups, with $N/2$ in each group. The experimental effect is often assessed using analysis of variance, but as Draper and Smith (1981, chap. 9) point out, any ANOVA model can be recast as a regression. Such recasting makes the following argument more transparent, so I adopt the regression representation of the experimental model. In this case, the regression model is straightforward:

$$Y = \alpha + \beta X + v$$

where Y is the response, X is the treatment dummy variable and v represents all other factors affecting Y.

It is amazing that we can estimate β without bias. After all, we are leaving out everything else in the world that affects Y and including only the experimental treatment, X, in the model. Consider the following example. We are interested in the effects of media bias on affect for public figures. We devise an experiment in which we expose subjects to either a positive message about the public figure or a negative one. But there are many other factors which influence affect, for example partisanship. Randomization in the experiment does not reduce the effect of partisanship one whit. So how can this possibly work?

Assume that the true model is

$$Y = \alpha + \beta X + \gamma Z + \epsilon, \tag{1}$$

where Z is partisanship, but what we estimate is

$$Y = \alpha + \beta X + v \tag{2}$$

where $v = \gamma Z + \epsilon$. In this case, X is fixed because it is under our experimental control. On the other hand, Z and ϵ are random variables outside of our control. Let us assume, as is usual, that these random variables are independent and identically distributed, so that $\mathrm{E}(Z_i) = \mu \, \forall \, i$, $\mathrm{E}(Z_i Z_j) = 0$, $\forall \, i \neq j$ and similarly for ϵ_i. We also assume that $\mathrm{E}(Z_i, \epsilon_i) = 0$. When we estimate β in equation 2 by OLS, omitting Z, we get

$$\begin{aligned}
\mathrm{E}(\hat{\beta}) &= \mathrm{E}\left[\frac{\mathrm{C}(X,Y)}{\mathrm{V}(X)}\right] \\
&= \mathrm{E}\left[\frac{\mathrm{C}(X,(\alpha+\beta X+v))}{\mathrm{V}(X)}\right] \\
&= \mathrm{E}\left[\frac{\mathrm{C}(X,(\alpha+\beta X)}{\mathrm{V}(X)} + \frac{\mathrm{C}(X,\gamma Z)}{\mathrm{V}(X)} + \frac{\mathrm{C}(X,\epsilon)}{\mathrm{V}(X)}\right]
\end{aligned}$$

where C and V represent covariance and variance, respectively.

The critical substantive observation at this point is that the random assignment of subjects to experimental groups guarantees that the expected covariances of X with Z and ϵ will be equal to zero. Once these terms are set to zero and so drop out of the expression above, it is easily seen to simplify to $\mathrm{E}(\hat{\beta}) = \beta$, so $\hat{\beta}$ is indeed unbiased.[1]

This is why experiments work: they avoid correlation of the treatment variable with all other influences. This means that the potential bias term in the expression above, $\mathrm{C}(X, \gamma Z)$ vanishes. In nonexperimental settings, such an assumption would be wild fantasy and we would be faced with a biased estimate. Instead, experimental control comes to the rescue. Notice, by the way, that the experimental randomization does not remove the effects of Z on Y. The effects of Z remain in the true model of the response, it is just that we can ignore it in estimating β.

When we estimate an experimental effect this way, we get unbiased estimates, but are we doing as well as we might? We should be troubled by the fact that we are throwing away relevant information about other influences on the dependent variable. It should never be inconsequential to discard relevant information. So what is the price we pay for the simplicity of the pure experimental effect model? Not bias, but efficiency.

The Price of Ignoring Information

Compare the variance for $\hat{\beta}$ estimated from equation (1) with the variance for $\hat{\beta}$ estimated using equation (2). The first variance is the standard result:

$$\mathrm{V}(\hat{\beta}_1) = \frac{\sigma^2_\epsilon}{N\mathrm{V}(X)(1 - r^2_{XZ})}. \tag{3}$$

[1] This substantive point is a little different from the subsequent technical proof, which I omit. A sketch of the proof is that since X is fixed, the expected values are taken for $Z - \overline{Z}$ and $\epsilon - \bar{\epsilon}$. These are then shown to be constants, so the covariance is seen to go to zero. This is a case, however, where I think the complete technical detail, while worth working through as an exercise, does not add to the substantive point. So I skip it.

Because it omits Z, equation 2 is a misspecified model but thanks to the experimental design, the estimator of β remains unbiased, as we have seen. The variance of this estimator is then

$$\begin{aligned} \mathrm{V}(\hat{\beta}_2) &= \mathrm{E}[\hat{\beta} - \mathrm{E}(\hat{\beta})]^2 \\ &= \mathrm{E}\left[\frac{\mathrm{C}(X, v)}{\mathrm{V}(X)}\right]^2 \\ &= \mathrm{E}\left[\frac{n^{-1}\sum(X_i - \overline{X})(\gamma(Z_i - \overline{Z}) + (\epsilon_i - \bar{\epsilon}))}{n^{-1}\sum(X_i - \overline{X})^2}\right]^2 \\ \mathrm{V}(\hat{\beta}_2) &= \frac{\sigma_\epsilon^2 + \gamma^2 S_Z^2}{N\mathrm{V}(X)} \qquad (4) \end{aligned}$$

where the last step depends on the assumption of independent observations, the independence of Z and ϵ, and that X is fixed.[2] Further, the variance of $Z, (S_Z^2)$, should be understood to be the sample variance of Z, rather than a population parameter.

Comparison of these two variances is revealing. Since the variance of Z must be non-negative, the numerator in equation 4 must be at least as large as that in equation 3. Which variance is larger depends on the correlation between X and Z, present in the denominator of equation 3. So long as $1 - r_{XZ}^2 > \sigma_\epsilon^2/(\sigma_\epsilon^2 + \gamma^2 S_Z^2)$, then the variance given by equation 4 will be larger than that of equation 3. But the experimental design virtually assures us of this. The random assignment of subjects to treatment and control groups means that the expected value of this correlation is zero. In actual experiments, there will be variation around this expectation. Nevertheless, it seems most probable that this correlation will be quite low in most experiments. Thus we can get more precise estimates of the experimental effects by including other influences in the model than we can by estimating the experimental influence alone.

Practical Implications

The practical implications of this are readily seen. For a fixed number of cases, we can get more precise estimates of the experimental effect by including additional relevant variables. Alternatively, for a fixed level of precision, we can reduce the number of cases needed. If there is a budget constraint, this means we can get equivalent statistical results for lower cost by introducing a more complete model specification.

These costs can be appreciated by comparing the number of cases required to make the two variances equal. Let N_1 be the number of cases used if only the experimental effect is included in the model, and N_2 be the number of cases when the partisanship variable is added to the specification. Then when these variances are equal we have

$$\frac{\sigma_\epsilon^2 + \gamma^2 S_Z^2}{N_1 \mathrm{V}(X)} = \frac{\sigma_\epsilon^2}{N_2 \mathrm{V}(X)(1 - r_{XZ}^2)}$$

We want to know how large N_1 has to be to equal the precision obtained with N_2 cases and a more complete specification. The answer is

$$N_1 = N_2(1 - r_{XZ}^2)\frac{\sigma_\epsilon^2 + \gamma^2 S_Z^2}{\sigma_\epsilon^2}$$

If the correlation between X and Z is close to zero, as is likely in an experiment, then the dominant effect is due to the relative sizes of σ_ϵ^2 and $\gamma^2 S_Z^2$. If the second of these is very small in relation to the first, then the gain in efficiency is minimal. However, if $\gamma^2 S_Z^2$ is close to the same size as σ_ϵ^2 then N_1 would be about twice as large as N_2, which is a very substantial cost.

This added expense can be avoided so long as it is cheap to acquire measures of additional relevant variables. In many experimental circumstances this should be easy to do, using a questionnaire, for example. Once these additional measures are at hand, it is trivial to estimate the more complete model in order to gain more precise estimates of the experimental effect. Thus the efficiency gains are both statistical and monetary.

An Example

As an example of these gains, I have created a Monte Carlo simulation of an experiment. In this experiment the true model is

$$Y = 1.0 + .65X + 1.0Z + \epsilon$$

where the variances of both Z and ϵ are equal to 1.0. This experiment is run with subjects randomly assigned to either experimental or control groups, with 25 subjects in each. The experiment is then replicated 500 times.

Table 1 gives the mean coefficients and mean estimated standard errors for two models, the first including only the experimental effect and the second also including the effects of Z.

Table 1: Comparison of Monte Carlo Results

Constant	1.0156	1.0153
	0.2820	0.2009
X	0.6389	0.6465
se(X)	0.3989	0.2845
Z	...	1.0018
se(Z)	...	0.1454

In this case, the estimated experimental effect is essentially the same for both models (.64 vs .65). However, the standard error for β is some 40% larger when the effects of Z are ignored (.40 vs .28). In this case, the average t-ratio

[2] Students should work out the intervening steps as an informative exercise.

is 1.60 when only the experimental influence is included, while it is 2.27 when Z is included. This shows that less efficient estimation may affect our substantive conclusions as well.

Conclusions

The aim of experiments is to test hypotheses. Most experimental analyses tend to focus only on the structure of the experiment while ignoring other influences on the dependent variable. This is valuable for simplicity and leads to unbiased results, as we have seen. This is a great strength of experiments. However, this common practice has two costs. First, by ignoring other influences on the response, the resulting model is a less complete substantive picture of behavior. Theoretical development might be enhanced by a more inclusive approach. Second, are costs in efficiency. Either greater precision can be achieved for a fixed cost, or equal precision reached for a lower price. Since resources are always limited, it makes sense to produce the most efficient estimates we can. Failure to do so may affect both our pocketbooks and our substantive conclusions.

References

Draper, Norman and Harry Smith. 1981. *Applied Regression Analysis, 2nd edition.* New York: Wiley-Interscience.

The Ills of Emphasizing Specification

Michael McDonald
SUNY—Binghamton

Charles Franklin recently remarked that there is a "tension between an interest in parameter estimation and a recognition of the uncertainties of model specification" (Franklin, 1990b: 1). I fear that political methodologists are attempting to alleviate that tension by putting second things first—i.e., estimation before inference.

As political methodologists have come to rely more and more extensively on econometric approaches to statistical methodology, the haunting specter of specification error—i.e., a mismatch between the true model and the estimated model—surrounds much that we do. It is ill-founded and can be ill-fated to think that the analyst should avoid specification errors, and thereby achieve what is termed pure-specification, by trying to verify that the estimated model is the true model. Such thinking is ill-founded because the avowed purpose of achieving the best estimation by avoiding specification errors supersedes the inferential purpose of observing relationships. It is ill-fated because, with a mistaken initiating premise, the methodological prescriptions it produces can lead to arguments that operate at cross-purposes. The conclusion to be reached is that worrying about arriving at the correct specification should be given a subsidiary role in research endeavors, but because it has been raised to a primary role it threatens to impede progress toward understanding.

To appreciate the distinction between treating specification as a primary versus subsidiary consideration, it is best to take one step back to the logical foundation of empirical inquiry. At its base, specification as a primary concern refuses to accept *modus tollens* $(P \Rightarrow Q; \neg Q; \text{therefore} \neg P)$ as the argument form of all theoretically grounded empirical inquiry. In words appropriate to empirical inquiry, *modus tollens* says:

> If my theoretical understanding (p) is true, then
> I should observe phenomenon q.
> I do not observe phenomenon q.
> Therefore, my theoretical understanding is not true.

Modus tollens implies a refutationist epistemology, to pursue understanding by testing propositions to see whether they are wrong. Instead of placing a primary emphasis on testing and inference, pure-specification emphasizes estimation. But, estimation without testing and inference is nothing more than description that may or may not be accurate. In place of refutation, the search for pure-specification tries to rely on verification. On this issue, however, we are often reminded, in more and less direct statements, that verificationism is impossible (e.g., Popper, 1959; Barry, 1970 [on the sociological approach]; King, 1989 [on inverse probability]).

These observations cannot be dismissed as fanciful epistemological musing. There are real, observable, and undesirable consequences for methodological practices derived from the elevation of pure-specification to a primary concern. Some of these are well known and often criticized, for example, maximizing R^2, forward stepwise regression, and similar variance explanation strategies. Others are more subtle. In the following paragraphs, I take examples from two prominent political methodologists to illustrate two less obvious consequences.

Estimation before Inference

Donald Green's (1990: 7-9) essay offers an opportunity to see how estimation can be given such importance as to replace testing and inference. Green created fictitious data sets, each with four variables, for analysis by approximately 200 first year Management students. His arrangement of this hypothetical world is expressed in two equations.

CONTRIBUTIONS = 15 + .15 PROFITS + .01 AGE + e1
CHAMBER RATINGS = 50 + 1.0 CONTRIBUTIONS + e2

Each student's task was to develop a regression model of charitable contributions by chemical corporations. Green created two difficulties for students: (1) to resist the temptation to include the endogenous RATINGS on the right hand side of their models and (2) to include the AGE variable which could be expected, given the error term, to be statistically significant only about 20% of the time. The students typically fell into both traps; only about a seventh of them reported the true model.

Green chastises the students for their typical practices of maximizing R^2 (the apparent motive behind including the endogenous RATINGS variable) and for the "equally prevalent and no less pernicious . . . tendency to drop variables that are deemed "statistically insignificant" " (Green, 1990: 9). There can be no defense for including the RATINGS variable, but the criticism for excluding this statistically insignificant variable has a ring of ill-logic to it.

Green clearly states that an analyst should not include just any insignificant variable. If there is theoretical reason to include it, however, then it ought to be included because exclusion "alters the sampling distribution of parameter estimates" (Green 1990: 8). This is not compelling. In the absence of knowledge of the true model, all it says is there are two sampling distributions between which the analyst must choose.[1] The first and clearest message to be appreciated by the analyst and reported to the reader is that the data are not up to the task of providing a strong test of the theoretical status of AGE and of providing as reliable an estimate of PROFITS as is desired. The reader deserves to hear that based on the evidence at hand there is no empirically justified reason for keeping AGE in the model. Failure to provide that inference depreciates the theoretical proposition to the level of dogma. It is futile to engage in empirical analysis and have readers read it when all one is allowed to learn is that *if* the theoretical proposition is correct, then the estimated coefficients are best and unbiased. The evidence supplies reason to believe the antecedent of that conditional statement is not true. That is a most important matter.

The problem that Green's data create emanates from the size of the standard error of the AGE variable. The only real solution can be obtained with a new design where AGE has greater variance, the N is increased, or both. With the new design and the new answer, we can address the question of whether the reliably estimated effect is large enough substantively to warrant our attention.

Looking for Verification and Finding Confusion

Larry Bartels (1990: 3) prefaces his remarks on five more or less able approaches to specification with the following observations.

> Statistical models are, and should be, more or less useful approximations. Nevertheless, the difference between more and less useful approximations to some underlying (and unknown) reality is a crucially important one.
>
> Second, I care fundamentally not about "explaining variance" or making forecasts but about estimating parameters of some theoretical interest.

It is easy to recognize, given his concern for estimating parameters of theoretical interest, that Bartels has in mind an approximation (i.e., a matching) of the estimated coefficients to the true coefficients. Unfortunately, without knowledge of true coefficient values, there is no obvious criteria by which to evaluate the match. Thus it becomes tempting to evaluate the approximation on the basis of the fit between the observed reality and what is predicted by the approximation. This, however, contradicts the indifference to explaining variance.

Bartels proceeds to offer harsh criticism of the stepwise regression approach to model specification. This computer assisted approach to finding a set of variables that produces a high R- square or adjusted R-square has the liability that the estimated coefficients "may bear little relationship to the structural parameter values of theoretical interest" (Bartels, 1990: 3). Curiously, however, Bartels finds qualified acceptance in evaluating a set of parameters through out-of-sample validation. The out-of-sample approach accepts previously estimated coefficients as given and applies them to new data. If the coefficients provide an accurate description of the new data, then he infers there ought to be a good fit between the true and estimated (from the original in-sample data) coefficients. The message is: do not get carried away with fitting the data at hand, but a good way to proceed is to fit the data not at hand. This recommendation, I submit, is confusing. Moreover, it cannot be salvaged by the verificationist hope that "in the long run, the right parameter values will tend to outperform the wrong parameter values when applied to new data" (Bartels, 1990: 4).

The two most obvious difficulties with Bartels' prescription are that: (1) a standard for evaluating the quality of the out-of- sample fit is absent, and (2) the coefficients may

[1] With Green's true model in hand, the desire to have sample estimate close to the true values, and the chance to look at only one sample, it is not entirely clear that one would want to include both PROFITS and AGE. The probability that the PROFITS coefficient is within some specified interval of .15 (e.g., .14 to .16) is virtually the same whether AGE is included or excluded. Also, when excluded, the AGE coefficient is certain to be understated by .01, but with AGE included there is a 50-50 chance of being even further away from the true value (i.e., below 0 or above .02). All in all, this is a nice example of a bias versus efficiency tradeoff.

bear little relationship to the structural parameters of interest. The plausible standard for judging an out-of-sample fit is the in- sample fit. Given that there is no way of justifying the statement that the in-sample fit is good or poor, there is not much to be learned by knowing the out-of-sample fit is just as good or poor. Second, wonderful out-of-sample predictions, even if they could be so denoted, can occur for a variety of reasons—spuriousness, indirectness, reciprocality, and complete reversal of causal order.

The approach to out-of-sample data analysis more in keeping with Bartels' prefatory principles and the *modus tollens* logical form is to: accept the originally estimated coefficients as expected values, acquire a new set of data, and estimate the same model. The evaluation is then based on matching the original and newly estimated coefficients. A mismatch tells one to rethink the theoretical proposition and derived model; an acceptable match says that the theoretical proposition still stands as a possibility that one can tentatively entertain a while longer. Concerns about spuriousness, indirectness, and various forms of endogeneity can be tested through appropriate adjustments to the model and subsequent re-estimation using the original and new data.[2]

Conclusion

The message here is really rather simple—there is not much future in conducting analyses guided primarily by a concern for the correct pure-specification. Rather, one must offer a conjecture, reason to its empirical implications, test whether those implications hold, and stick with the conjecture tenaciously until it is demonstrably refuted. With pure-specification as our primary guide we are led to non-cumulable empirical results.

Nothing I have said is a warrant to forget about pure-specification errors and their implications. We need to study them intently and intensively. Pure-specification errors inform us about a variety of ways in which our theoretical propositions may fail. Generally they remind us that "[t]he stochastic component is not a technical annoyance, as it is sometimes treated, but is instead a critical part of the theoretical model" (King, 1989: 9). Observing these errors is, perhaps counter intuitively, a sign of progress. They provide a refutation of our thinking and thereby knowledge we had not had before. What is more, the observed error points us toward probable sources of erroneous thinking—e.g., autocorrelated errors suggest an incorrect functional form, or a missing variable, or an incorrectly incorporated lag structure; heteroscedasticity indicates the possibility of a missing interactive variable; the inclusion of a seemingly irrelevant variable whose coefficient has a t-value greater than 1.0 provides reason to think there is systematic variance in e_i beyond what the theoretical proposition deems relevant; a stochastic component in X indicates a variance in e_i that is being attributed to Y whereas its measurement belongs to X and can be eliminated with a more reliable measurement of X.

The recommendations are:

1. Work fervidly to translate words to equations. Often this is not intuitively obvious (e.g., see Franklin, 1990a).
2. When words are too vague to provide firm expectations, take several possible interpretations and test them as alternatives.
3. Evaluate the parameter estimates for their fit to the expected values of the coefficients.
4. Test whether the evaluation in point 3 is adversely affected by pure-specification errors. If so, then in sequence:
   (a) rethink the theoretical proposition,
   (b) rethink the design ("design is data discipline," as Kerlinger [1986:302] tells us; political methodology appears to have forgotten that as it has moved toward econometric methods), and
   (c) as a last resort, try to find an appropriate technical adjustment.
5. Accept the *modus tollens* argument form and its refutationist method as the logical undergirding of theoretically inspired empirical inquiry; it is all we have.

References

Barry, Brian. 1988. *Sociologists, Economists and Democracy*, Midway Reprint Edition. Chicago: Univ. of Chicago Press.

Bartels, Larry M. 1990. "Five Approaches to Model Specification." *The Political Methodologist* vol. 3, no. 2:2-6.

Franklin, Charles H. 1990a. "More Words and a Picture about Words and Pictures." *The Political Methodologist* vol. 3, no. 1:13-14.

Franklin, Charles H. 1990b. "Notes from the Editor." *The Political Methodologist* vol. 3, no. 2:1-2.

Green, Donald P. 1990. "On the Value of Not Teaching Students to be Dangerous." *The Political Methodologist* vol. 3, no. 2:7-9.

Kerlinger, Fred N. 1986. *Foundations of Behavioral Research*, 3rd edition. New York: Holt, Rinehart & Winston.

[2] Bartels out-of-sample validation example concerns an instance where only three new data points have become available. Data limitations of this sort do warrant compromises. Nevertheless, they ought to be seen as compromises akin to those made when theoretical limitations lead one to such compromises as building a model on goodness-of-fit criteria. In both cases, the analysis provides new information, but skepticism is to be advised.

King, Gary 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* New York: Cambridge.

Popper, Karl 1959. *The Logic of Scientific Discovery.* New York: Harper & Row.

Further Thoughts on Data-Dredging: A Reply to McDonald

Donald P. Green
Yale University

Model specification rightly occupies a central place among the concerns of political methodologists. Perhaps this would not be the case were political science more reliant on controlled experimentation. But the fact that political scientists subsist on quasi-experimental data means that specification assumptions play a decisive role in shaping statistical inference. How stable are mass policy attitudes over time? The answer hinges on what one assumes about measurement error (Converse, 1964; Achen 1975). To what extent does incumbent spending influence House election outcomes? The answer hinges on one's perspective on the problem of simultaneity (Jacobson 1978; Green and Krasno 1988). How does economic performance influence presidential popularity? The answer hinges on what one assumes about serial correlation in the disturbance term (Mueller 1970; Hibbs 1973-74). Thus, it is hardly the case that political methodologists such as myself are more interested in specification than inference. The fact of the matter is that problems of inference in political science tend to be embedded in problems of specification.

Where Michael McDonald and I disagree is not over the primacy of specification vis-à-vis inference. Rather, we part company over how specification is best achieved so as to support inference.[1] My position is that regression diagnostics such as R^2 or t-ratios are generally unreliable guides to model specification and should be eschewed in favor of theory-based specification decisions. McDonald's view is less clear-cut. On the one hand, he indicates that a significant t-ratio is "no defense" for including a regressor (RATINGS) that arguably should be excluded on theoretical grounds. So far, so good. But then McDonald reverses course and asserts that variables may be dropped when "there is no empirically justified reason" for keeping them in the model. My criticism of the practice of dropping statistically insignificant variables, in McDonald's view, "has a ring of ill-logic to it."

It is unclear from the prescriptions McDonald makes how he integrates these two decision rules. I gather that his policy for model specification is something like this: variables are to be excluded if they fail to meet two conditions, (i) they must be theoretically defensible and (ii) they must achieve statistical significance. RATINGS fails on (i), and AGE fails on (ii). I do not regard this as a sound policy.

A well-known econometric result shows that when deciding whether or not to include a potentially irrelevant regressor (i.e., a regressor that may have a structural effect of zero), one should consider whether the variable's true effect is likely to exceed the true standard error associated with the regression estimate. If the effect is thought to be larger, then the variable should be included. If not, the variable should be excluded.[2] With regard to my simulation, this implies that the mean squared error associated with the estimated effect of RATINGS will be smaller if we exclude AGE, provided that AGE's true effect is smaller than its true standard error.

Two points should be noted in light of this result. Inconsistent with the MSE result is the notion that variables which fail to achieve conventional levels of statistical significance (e.g., alpha=.05) should be excluded. This practice is likely to lead to biases that are not justified by savings in efficiency. Hence, my students were mistaken to operate as though regressors must prove themselves worthy of inclusion into a regression equation by being statistically distinguishable from zero. Second, the MSE result does not involve sample estimates; instead, it has to do with the true parameter and its true sampling distribution. In constructing the simulation, I tried to make the theoretical status of AGE somewhat ambiguous and deliberately set the ratio of these two quantities close to one, so that it would be difficult for students to form a firm opinion about the true ratio based on introspection or inspection of the data. I would hardly be justified in chastising students for dropping AGE if their decisions to do so were based on their priors concerning this ratio.[3] But it is apparent from the pattern of results that students were interested in whether AGE was statistically significant, not whether its true effect exceeded the estimate's true standard error.

McDonald seems to suggest that readers have a right to know about the sensitivity of the estimates to different specification assumptions. I agree. If several plausible specifications exist, one reasonable strategy is to present the reader with a set of regression results, so as to illustrate the robustness of the results. But this objective is

[1] McDonald is more convinced than I of the philosophical merits of falsificationism, but this brief reply is not the place to take up a topic of such complexity.

[2] See Johnston (1984: 253-259) for a discussion of MSE as a standard by which to evaluate bias–efficiency trade-offs.

[3] Incidentally, although I heaped scorn on the general practice of dropping insignificant variables from regression equations, I was rather lenient on those who elected to drop AGE, since I knew its presence or absence did not have profound effects on the influence of PROFITS and that students might have different priors about whether it belonged in the model.

not achieved by discarding regressors when they turn out to be statistically insignificant. Furthermore, it should be noted that in the case of irrelevant regressors, such sensitivity analyses may be a form of overkill. What is so bad about presenting an equation with a statistically insignificant parameter estimate? At worst, the sampling variability of the estimates will increase. Weighed against the alternative error–excluding a relevant regressor–the consequences of including an irrelevant regressor seem pretty tame.

McDonald is correct to point out that political scientists may at times find themselves in the position of saying that the substantive inferences they draw from their regression results hold provided their specification assumptions are correct. He bristles at the notion that inference could rest on "dogmatic" specification assumptions, but I see this as an unavoidable by-product of the uncertainty surrounding quasi-experimentation. Many debates in political science are currently in this state of theoretical deadlock. But rather than try to data-dredge our way to inference, it might be more fruitful to devise experiments or quasi-experiments that speak to the most contentious specification issues and reassess our quasi-experimental results in light of these findings.

References

Achen, Christopher H. 1975. "Mass Political Attitudes and the Survey Response." *American Political Science Review* 69: 1218-31.

Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics." In *Ideology and Discontent*, ed. D. Apter. New York: Free Press.

Green, Donald P. and Jonathan S. Krasno. 1988. "Salvation for the Spendthrift Incumbent: Reestimating the Effects of Campaign Spending in House Elections." *American Journal of Political Science* 32: 884-907.

Hibbs, Douglas A., Jr. 1973-74. "Problems of Statistical Estimation and Causal Inference in Time-Series Regression Models." In *Sociological Methodology*, ed. H. Costner. pp. 252-308.

Jacobson, Gary C.. 1978. "The Effects of Campaign Spending in Congressional Elections." *American Political Science Review* 72: 469-91.

Johnston, J. 1984. *Econometric Methods*. 3rd ed. New York: McGraw-Hill.

Mueller, John H. 1970. "Presidential Popularity from Truman to Johnson." *American Political Science Review* 64: 18-34.

Goodness of Fit and Model Specification

Michael S. Lewis-Beck
University of Iowa

Andrew Skalaban
University of California, Davis

On certain subjects there can never be a "last word." However, on the subject of R-squared this is our last (?) word. (For earlier efforts, see Lewis-Beck and Skalaban, 1990a, 1990b). Here we focus on the meaning of R-squared when the issue of model specification is raised. Of course, with a misspecified model the R-squared is of no value, and neither are the regression coefficients nor the tests of significance. However, under the assumption of a correctly specified model the R-squared can be quite useful. Below, we outline these interpretations, then address the question of the R-squared in practice. By teaching students to employ R-squared in political research do we encourage bad methodological habits? Maybe for some. But we believe the good from the many outweighs the bad from a few.

Strong Specification Assumption

A strong specification assumption means that the lucky researcher has correctly identified his or her regression model; namely, all and only the appropriate variables are included. Under such conditions do we care about the information contained in R-squared? Some political scientists may not, preferring only to concern themselves with the parameter estimates for the intercept and slopes. After all, once the model is correctly identified, any unexplained variation is random. Should we care about randomness, once we have explained all that can be explained? Yes. Knowing the proportion of the variation in the dependent variable that is truly random is an inherently interesting question. (We admit our bias. The "free-will" versus "determinism" controversy fascinates us, as it has the philosophers.) Furthermore, the question of randomness should capture the more earthly attentions of those modelers wishing to make predictions.

Under the strong specification assumption, the R-squared provides a consistent parameter estimate of the proportion of non-random variation in the model. As well, it measures relative linearity (the base being perfect linearity between the Xs and Y). As Blalock (1960, p. 311) once observed, in "many practical sociological problems the linear models " is "close enough." Indeed, in the study of political problems, we find over and over again that the linear model is hard to improve upon. And, even when departures from it are necessary, linearity provides and important, universal base for comparison.

Neither of these pieces of information — non-random variation or linearity – are directly available from the Standard Error of Estimate (SEE). Does this mean that the R-squared is the preferred goodness-of-fit statistic? Only if those are the pieces of information you want. If you want to see how close the model's predictions are to zero, then consult SEE as well. Does this mean that a model with a high R-squared is a "better" model than one with a low R-squared? Certainly not, if both are correctly specified. It merely means that the first model has more structure, less randomness, than the second. But, if we had to bet our $64 on the accuracy of prediction from one of two correctly specified models with different dependent variables, then we would choose the one with less random variation— the one with the higher R-squared.

Weak Specification Assumption

The weak specification assumption is met when the researcher has a realistic confidence about the basic theory embodied in the model, and has committed none of the mortal sins of regression analysis; for example, no endogenous variables are included as predictors, no clear hierarchy of causation exists among the independent variables, the model is not "overfitted". Under such conditions looking at goodness of fit, especially the R-squared, can be an aid to model building.

Suppose we compare two models of the same dependent variable, both of which meet the weak specification assumption. Then, the R-squared can provide a useful tool in deciding whether or not to add an additional independent variable. But, the critic may retort, why not simply rely on the t-test of the coefficient? Unfortunately, the world is sometimes too complex for that simple rule. Think about this scenario. You add a new independent variable to a model with two other predictors. The new variable is somewhat, but not highly, collinear with one of the original, highly significant predictors. When all three are included in the regression, the new variable turns out to be marginally significant. But the problem is that the other variable with which it is correlated is now no longer highly significant, but is itself only marginally so. Do you include the new predictor? The pat answer is to rely on theory, but theory too can be ambiguous.

One way to think about the problem is to consider whether the new model better accounts for the variation in the dependent variable. What constitutes significantly more explained variance is, in the end, a matter of judgment. You can do formal tests on the two R-squared or error sum of squares (it amounts to the same thing). Or, you can set some substantive level of increment to the adjusted R-squared that you think justifies changing your original model. Of course, that justification must have a strong theoretical component; otherwise, it is simply mindless R-square maximizing.

R-Squared in Practice

Many political methodologists have warned against the abuse of R-squared as a measure of a regression model's worth. We support this warning. (Contrary to popular belief, we are not writing a book entitled *The Joy of R-Squared*. However, some would take the argument against the R-squared further than we. The general tone of the critics is that "use invites abuse". In particular, it invites comparison when comparisons are neither appropriate nor relevant (read King 1986; 1990). It can be a disaster for model identification because it tempts one to include endogenous variables on the right-hand-side (read Green, 1990). It too easily serves as a substitute for other statistics, such as the SEE, that contain more important information (read Achen 1982). It doesn't measure anything we are much interested in anyway (read all of above).

For some time, we have been pushing R-squared as good medicine. Is the dosage reaching a toxic level? Should we "just say no" to R-squared? We would rather not. Maybe we are hooked, but it is a statistic that we turn to repeatedly. For example, when evaluating model with thousands of cases, so that nearly everything is significant. Or, when reading the literature, and trying to figure out the metric needed to interpret an author's reported SEE. Or, when trying to judge how well all those independent variables account for what there is to account for. Under these, and other circumstances we have documented elsewhere, R-squared can be a vital piece of information.

References

Achen, Christopher H. 1982. *Interpreting and Using Regression*. Beverly Hills: Sage.

Blalock, Hubert M. 1960. *Social Statistics*. New York: McGraw-Hill.

Green, Donald P. 1990. "On the Value of Not Teaching Students to be Dangerous". *The Political Methodologist*, vol. 3, no. 2, 7–9.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science". *American Journal of Political Science*, 30: 666–687.

King, Gary, 1990. "When Not to Use R-Squared." *The Political Methodologist*, vol. 3, no. 2, 11–12.

Lewis-Beck, Michael S. and Andrew Skalaban. 1990a. "The R-Squared: Some Straight Talk." *Political Analysis*, 2: 153–172.

Lewis-Beck, Michael S. and Andrew Skalaban. 1990b. "When to Use R-Squared." *The Political Methodologist*, vol. 3, no. 2, 9–11.

R-Square Encore

Robert C. Luskin
University of Texas

Some time ago, I served as discussant on a panel on which Mike Lewis-Beck and Andrew Skalaban were presenting a paper in defense of the coefficient of determination or R^2, an unpopular statistic in methodological circles these days (Achen, 1982; King, 1986) The Lewis-Beck/Skalaban defense was spirited, and I commented at the time that my own attitude, embodied in a paper I had just at that point shipped off to the *American Journal of Political Science*, was more "Minnesotan"–"not so bad, could be worse." Although my paper, now forthcoming in *AJPS* (Luskin, 1991), is concerned with a cluster of related statistics—bivariate (Pearsonian) correlations and standardized regression coefficients as well as R^2s—my mandate here is confined to R^2.

Few statistics are more widely used, or abused. Modelers compare R^2s from radically different models, crow over R^2s whose elevation inheres in the model or data, and worst of all toss theory to the winds in the single-minded pursuit of higher R^2s. The gist of the relevant portions of the paper I have been referring to, however–and the argument I wish to make here–is that the R^2 is not therefore useless. *Abusus*, as the Latin tag in the paper's title has it, *non tollit usum*.

R^2 as a Measure of Fit

The controversy over R^2 revolves around two issues, seldom if ever well separated. The first concerns R^2 as a measure of fit. Everyone agrees that R^2 measures the sample regression function's ability to reproduce the sample data. [1] King (1990) argues that R^2 is inferior, even for this purpose, to $\hat{\sigma}_u$, the estimated standard deviation of the disturbance u, because the latter is in the same units as the dependent variable y. I should contend, to the contrary, that the only way we can tell whether the variation expressed by $\hat{\sigma}_u$ is large or small is with reference to the variation in y, a comparison merely formalized in R^2.

But the choice between R^2 and $\hat{\sigma}_u$ is relatively small beer. The main point of contention under this heading is whether either of these statistics says anything about the truth of the model. Here we come rapidly into deep water. What *is* the truth we are trying to model? Is the true model inherently stochastic or deterministic? (See King, 1990.) *Is there*, indeed, a true model? I should rather suggest that

> Uniquely true models exist only in the assumptions of econometric proofs. A given y can always be explained in a number of equally valid ways–in terms of a larger set of conceptually finer x's or a smaller set of conceptually grosser ones, in terms of variables that have their effect at close quarters or variables that act from afar. At most, there may plausibly be a single true model *of a given type*–at a given level of conceptual aggregation, at a given causal distance–and this plurality of standards complicates evaluations and comparisons (Luskin, 1991).

Partly because it is what I believe and partly because it is more problematic for R^2, we may take the true model of a given type to be stochastic, with a disturbance of unknown but nonzero variance $\sigma_*^2 \leq \sigma_u^2$. Now much depends on our faith in our model. If the model is gospel, σ_u^2 $(= \sigma_*^2)$ is what it is, period; if σ_u^2 is large, y is just poorly explicable. If, more realistically, however, the model may not be true, a large σ_u^2 tells against it.[2] How large is σ_u^2? The natural measure is $P^2 = 1 - (\sigma_u^2/\sigma_y^2)$, the population analog of R^2. On the provisional assumption that the regressors are uncorrelated with u,$0 \leq \sigma_u^2 \leq \sigma_y^2$, and consequently $0 \leq P^2 \leq 1$.

Of course, we cannot actually determine P^2, but estimating it brings us back to R^2. Contrary to the critical line about R^2, but consistent with a number of treatments in statistics and econometrics, R^2 is a consistent though biased estimator of P^2. The bias can be large when the number of regressors is sufficiently large in relation to the number of observations and P^2 is sufficiently small. These conditions are fairly rare, however, and the *adjusted* R^2

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-K-1}(1-R^2)$$

is virtually unbiased even then (see Montgomery and Morrison, 1973).[3]

A greater problem is that the P^2 for the true model is likewise unknown, which in turn keeps us from knowing exactly how high a P^2 we should be satisfied with (as King, 1990, points out). Yet this does not void the P^2, or its estimator R^2, of meaning completely. One would have to take an awfully dim view of the fathomability of human behavior to regard P^2s of say .2 as reason for contentment. Beyond that, the modeler can draw on his or her knowledge of how causally close y is to the regressors, how conceptually aggregated and how well measured the variables are, and how aggregated the units of analysis are. What is high and what

[1] I shall assume for simplicity's sake that we are dealing with missionary style regression analysis—a linear model estimated by OLS—although a fair proportion of these remarks would carry over to more complex cases. For a discussion of the definition and interpretation of R^2 outside OLS, see Luskin (1984).

[2] It is important here to be clear about the distinction between the *sample* variance $s_{\hat{u}}^2$ of the *residual*, $\hat{u}$, and the *population* variance σ_u^2 of the disturbance u. Minimum σ_u^2 is a reasonable definition of the true model; minimum $s_{\hat{u}}^2$ is not, and is not even an especially reliable guide to minimum σ_u^2.

[3] I shall continue referring simply to R^2, even though R^2_{adj} is clearly the preferable statistic (by a small margin in most cases, but by a wide margin in some).

is low will depend on the nature of the variables, the data, and the model. But the lower the P^2, the higher, *ceteris paribus*, one's "specification anxiety" should be. And an R^2, estimating P^2, that is too low should be an occasion to think some more, and perhaps to remodel.[4]

The Trouble with Fit

The second issue concerns fit as a criterion for model selection. The chief complaint on this count is well captured by the story told in the midwest of the farmer who goes to town to buy a new Sunday suit. The salesman in the haberdashery takes one look and decides that this is the moment to unload that oddly cut plaid on the back rack. The farmer is perceptive enough to see, when he tries it on, that the suit does not actually fit very well, but the salesman assures him that on the contrary it fits very well indeed—all he has to do is to lower his left shoulder, bend forward a little bit, keep his right knee slightly flexed, and crane his neck a bit to the left. The farmer is convinced, buys the suit, and leaves the store wearing it, carrying himself as instructed. As he walks out into the street, one passerby says to another, "Look at that poor, unfortunate man. It must be terrible to have to go through life so deformed." "Yes," says his companion, shaking his head sadly, "how true ... but what a great fitting suit!"

The moral is never to deform one's theory to achieve better fit— never, among other things, to set out simply to maximize R^2. But this is not really a story about R^2, as opposed to other criteria of fit; it is a story about specification searches insufficiently channeled by theory. The danger, as King observes, is "overfitting." Beyond a point, the model can only fit the sample better by fitting the population worse (by being less true, in the sense described above). But the same danger lies in giving up the reins to any measure of fit. Shaping the model to maximize the significance of the F test or the number of significant t tests is similarly misguided.[5]

R^2 and the Art of Statistical Analysis

The critics of R^2 tend to look more kindly on statistical tests, the most directly related of which is the "F test" of the null hypothesis that all the regression parameters, save the intercept, are zero–that none of the regressors actually has any effect. Certainly we should like to reject this. But how content should we necessarily be, having done so? The manufactured but plausible example in Luskin (1991) has $n = 2808$, $K = 7$, and $R^2 = .1$. The F statistic, at 44.4, falls clear off the table. We can be highly confident that the model explains *something* of y. *How much* does it seem to explain? Very little. The R^2, estimating P^2, is only .1. The model is statistically but not substantively significant.

The difficulty with the notion of substantive significance, of course, is that it lacks the surface precision of the probabilities in statistical tests. Measures of substantive significance require artful and to some extent subjective interpretation, and I suspect that it is uncomfortableness with this need for tact that underlies many of the objections to R^2. This is a fastidiousness I think should be overcome. To be assured that one's model is worth something, however little, is not enough; we also need to address the question of how worthy it is.

It should be clear that R^2 as a topic opens out onto a variety of other topics having to do with model assessment and specification searches, the largest question about which is how we can best grope our way asymptotically toward the truth. Recent years have seen the introduction of a variety of new devices . There are measures of fit like Mallows' C_p, Amemiya's PC, and Akaike's AIC (all relatively simple functions of R^2, adjusted like R^2_{adj} for degrees of freedom consumed), and there are specification tests, like Ramsey's RESET test, the Hausman test, the Cox-Peseran test, and the J test.[6] And some of these devices may well do better at steering us toward the true model than R^2 or R^2_{adj}. Other useful approaches include exposing one's model to fresh data, putting it through a "sensitivity analysis" to see how critically the results depend on the various details of model and assumptions, and incorporating prior information, à la Bayes or otherwise (all sketched, with customary elegance, by Bartels, 1990).

As advertised, then, this is only a Minnesotan defense. R^2 is only roughly informative. Other statistics may be more helpful in specification searches. And neither R^2 nor any other statistic should be allowed to dictate the model. Yet it is still a defense, against criticisms that are often too harsh. R^2 *is* at least roughly informative, even as regards specification.

References

Achen, Christopher. 1982. *Interpreting and Using Regression.* Beverly Hills: Sage.

Amemiya, Takeshi. 1980. "Selection of Regressors." *International Economic Review,* 21: 331-354.

Bartels, Larry M. 1990. "Five Approaches to Model Specification." *The Political Methodologist* vol. 3, no. 2:2-6.

Fomby, Thomas B., R. Carter Hill, and Stanley R. Johnson. 1984. *Advanced Econometric Methods.* New York:

[4] Note, by the way, that this logic involves no appeal to "inverse probabilities" (cf. King, 1990). Rather, we assume the model is correct and see what estimate of P^2 that leads to.

[5] Note too that the recycling of data in specification searches, however conducted, results in a *pretest estimator*, whose true statistical significance is far lower than the p-values on the printout suggest (Leamer, 1978, pp. 88-89; Lovell, 1983).

[6] For more details, see Fomby, Hill, and Johnson (1984, pp. 400-36), and Judge et al. (1988, pp. 838-58).

Springer-Verlag.

Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lutkepohl and Tsoung-Chao Lee. 1988. *Introduction to the Theory and Practice of Econometrics,* (2nd edition). New York: Wiley.

King, Gary. 1986. "How not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science,* 30: 666-687.

King, Gary. 1990. "Stochastic Variation: A Comment on Lewis-Beck and Skalaban's 'The R-Squared'." *Political Analysis,* 2: *in press.*

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley.

Lewis-Beck, Michael S., and Andrew Skalaban. 1991. "The R-Squared: Some Straight Talk". *Political Analysis,* 2: *in press.*

Lovell, Michael C. 1983. "Data Mining". *Review of Economics and Statistics,* 65:1-12.

Luskin, Robert C. 1984. "Looking for R^2: Measuring Explanation Outside OLS." *Political Methodology,* 10: 513-532.

Luskin, Robert C. 1991. "Abusus Non Tollit Usum: Bounded Truthfullness in Statistics." *American Journal of Political Science,* 35: in press.

Montgomery, David B., and Donald G. Morrison. 1973. "A Note on Adjusting R^2." *Journal of Finance,* 28:1009-1013.

Poisson Models of Elite Turnover: Two Memoirs

Thomas W. Casstevens
Oakland University

W. Allen Wallis
American Enterprise Institute

Casstevens:

In 1968-69, I read mathematics at Dartmouth College as a National Science Foundation Science Faculty Fellow. The winter months saw me reading probability theory with J. Laurie Snell. I mastered *inter alia* chapter seven, particularly the negative exponential of radioactive decay, in his co-authored textbook, *Finite Mathematical Structures* (1959). And I browsed, especially in Emmanuel Parzen, *Modern Probability Theory and Its Applications* (1960).

I was struck by two things: the downward deep of the decay curve and Parzen's summary of W. Allen Wallis, "The Poisson Distribution and the Supreme Court," *Journal of the American Statistical Association,* 31:376-380 (1936).

For weeks, the decay curve was drawn on the blackboard in my office in Bradley Hall. I often stared at it. "I know something that looks like that" was my recurring thought. Finally, one day, came the stray thought, "of course, legislative turnover looks like that." (I should have thought "tenure" but, in fact, "turnover" was the word that came to mind.) I now believe that thought was derived unconsciously from a graph of tenure in J.F.S. Ross, *Parliamentary Representation* (1948). Ross' book was a source for my Senior Thesis at Reed College (1959).

Wallis' article is a beautiful application of the Poisson distribution to the turnover of justices on the Supreme Court. I thought it was quite general in its application to political elites. Within days, a replication cum extension was completed, "Poisson explanations of turnover in the British offices of the prime minister and the monarch" (1969). But that was not published. Editors of journals of political science seemed to view it as trivial and not true. Nevertheless I persisted.

I did not see for a year, or thereabouts, that the two models are essentially equivalent: Poisson turnover implies exponential tenure, and vice versa, under broad conditions. The rate of turnover and the length of tenure are multiplicative inverses.

Since those days, this model of turnover and tenure has flourished in political science. I recently reviewed the literature and concluded simply: "The source of this model of the circulation of elites is Wallis (1936). The model has been rediscovered from time to time; it is time for us to make it a permanent part of our standard repertoire. This is an idea whose time has come." (*American Journal of Political Science,* 33:294-317 (1989). A copy of that review was sent to W. Allen Wallis. His reply follows.

Wallis:

My thanks indeed for your article on "The Circulation of Elites" in the *American Journal of Political Science* for February.

The idea that my 1936 article initiated an interesting line of research comes as a shock—a most pleasant shock, of course. I found your article quite interesting, not least the extensive bibliography.

My paper on the Supreme Court was written originally as a term-paper for a graduate statistics course at Columbia with Harold Hotelling, under whom I studied in 1935-36. That was the period when the Supreme Court was overturning key "New Deal" laws, most conspicuously the NRA and the AAA. Much comment at the time had it that the administration had erred in postponing adjudication as long as possible, presumably in the expectation that the NRA and AAA would be so interwoven into the fabric of the economy, and so popular, that the Court would not upset them.

It occurred to me that the strategy might have had a different basis, the expectation that as time passed the composition of the Court would change in the Administration's favor. This raised the question, if there was a three year delay (as there was) what was the chance of getting at least one new Justice? As it turned out, of course, the NRA lost unanimously and the AAA by at least two (if I recall correctly, three) votes; but no doubt the Administration expected to do better.

Incidentally, the article contains an error in the chi-square test of the fit of the Poisson to the actual data. To determine the probability corresponding with the value of chi-square, I used the Wilson-Hilferty normalization. I forgot, however, that this gives a two- tail probability where one-tail is appropriate, and is, therefore, double the correct probability. I reported the probability to be something like 0.94 (I do not have a copy of my article here), almost too good a fit. The correct probability, 0.47, strengthens my claim for the appropriateness of the Poisson distribution.

My work with the Poisson distribution turned out to be invaluable to me during the Second World War, when I worked on military problems. In many studies of hits on targets, of equipment defects, etc., the Poisson distribution or the related negative exponential are central.

Many thanks for your thoughtfulness in sending me your article. My father, an anthropologist, had a somewhat similar experience when an article he had published in the mid-teens was republished at Berkeley in the mid-fifties, forty years later. He was not sent a copy of the reprint because it was assumed that he was long since dead and gone. Your article cited mine fifty-three years later, and I am happy that you knew my whereabouts and took the trouble to send me a copy.

Review of *Markov*, a New Gauss Program

Nathaniel Beck
University of California, San Diego

I just received a commercial beta test copy of *Markov*, the new "front-end" for *Gauss* from Aptech Systems. *Markov* requires *Gauss* 2.1; like *Gauss*, it can run on a simple 8086 system, but for performance it requires a 386 so that *Gauss*/386 can be used. It should be commercially available by the time you read this. *Markov* was written by Scott Long of the Sociology Department at Indiana University. He is the author of Aptech's limited dependent variable set of programs, and these programs are the heart of *Markov*.

All of us who use *Gauss* invariably wish for a nice front-end. *Markov* certainly goes a good way in this direction. It provides commands to easily look at the contents of data matrices, get variable names and edit matrices. (Actually *Gauss* 2.1 is much friendlier in this regard, giving an (optional, at extra cost) full screen editor for entering matrices. It is this editor which is used by *Markov*.) The ability to more easily manage data and keep track of data sets will make *Markov* useful to all *Gauss*ians.

Markov contains a module for data selection and transformation. This process is friendly but limited. It does not have the power of SST or Limdep or SAS, but it may simplify the lives of students. But for those who want more flexibility you still have all the *Gauss* commands. This is the beauty of *Markov* - if you can't do it in *Markov* you just do it in *Gauss*, with all the incredible flexibility offered by *Gauss*. *Markov* runs under *Gauss*, and, except for some use of memory, imposes almost no costs on the user. (Memory loss should be relatively unimportant for 386 uses, but it may cause problems for 086 users.)

Markov contains modules for regressions, limited dependent variable analysis and log-linear analysis, as well as for standard descriptive analysis. The modules are very good although they don't contain all the bells and whistles of Limdep or SST. The regression module will do 95% of what most people do, giving you heteroskedastic consistent standard errors, the Belsley, Kuh and Welsch set of diagnostics, WLS and 2SLS/3SLS/SUR.

But the wonderful thing is that if *Markov*'s regression package doesn't do what you want, once again *Gauss* saves the day. *Markov* allows access to all relevant matrices. So if you want to, say, used jackknifed standard errors, you can start with the regression routine in *Markov* and then do the rest of the computations in *Gauss*. While Limdep, SST and SAS also allow matrix manipulation, none of them have *Gauss*'s power and flexibility.

The same thing is true for the limited dependent variable module. It covers the garden variety logit/probit models which meet a huge proportion of most users needs. It doesn't have the limited dependent variables bells and whistles of Limdep, but if you want to do state of the art work, once again *Gauss* is sitting there, waiting to serve.

The graphics in *Markov* is primitive, but *Gauss* provides fabulous professional quality graphics (much better than any of the standard PC statistical packages). So you run your regression in *Markov* and plot the residuals with one call to *Gauss*. Not a bad compromise at all. (*Markov* itself will probably get a better set of graphics, which will make it easier to use the powerful *Gauss* graphics.

If you think of *Markov* purely as a statistical package, it is on the low end, with much less power than SST or Limdep (or SAS). But even as a package it does 95% of what most of us do every day (and is quite adequate for a graduate student lab). But the real power of *Markov* is the ability to do most of the things we do easily while retaining the power of *Gauss* when that power is needed.

Markov is also very nice to use in a graduate student lab for any methods course at the level of, say, Hanushek and Jackson, or above. I like my students to both do OLS on real data and also to understand the matrix algebra behind the package results. With *Markov* you can do both, and without the steep learning curve that goes with *Gauss*. Many students will never go beyond *Markov* but those with greater needs will find themselves having a much easier time with *Gauss* later on. If you just use *Markov* as a lab package you can probably do better with some other package; it is the flexibility that *Markov* offers that makes it worthwhile. Of particular interest to instructors is the ability to write your own routines and modules; a customized package for instruction (and research) becomes a real possibility (but creating such a customized package is not cost-free). For example, it should be possible to hand Gary King's Count programs off *Markov* relatively easily.

In short, if you just want a single package (either for instruction or research) and you don't want the extra flexibility afforded by *Gauss* then I don't think I would choose *Markov*; if you want a package that will allow students to grow into maximum likelihood, *Markov* really has no competitors. If you always wanted to teach with *Gauss*, but couldn't deal with the steep learning curve, then *Markov* may be just the thing.

If you like *Gauss*, at worst *Markov* can't hurt. At 3AM its friendlier data handling, data archiving and simple reading and writing commands should be a real joy. But if you are one of the people who say "I really should look at *Gauss*" or "I looked at *Gauss*, it looks great, I even own it, and one day I will get around to figuring it out," *Markov* may be the perfect tool for you.

Markov is available from Aptech Systems, 26250 196th Place SE, Kent, Washington, 98042, (206) 631-6679. *Markov* requires *Gauss* 2.1 which costs $495 for the standard version and $695 for the 386 version (which I recommend for research purposes). *Markov* costs $195, plus extra for the matrix editor ($40). Aptech has a very generous policy about site licenses and inexpensive student versions.

The *Gauss* Corner: A Stochastic Censoring Tobit Model

While *Gauss* may be the most powerful statistical programming language around, it is also the most infuriating, with error messages designed to drive one to either an asylum or a monastery for mental or spiritual relief. As a modest effort toward the mental health of the discipline, *TPM* will occasionally publish *Gauss* examples as a means of furthering the use of the software and of exposing a wider audience to its charms.

The following *Gauss* program computes a stochastic censoring tobit model, a highly useful technique not available in standard statistical packages. The model is described in Amemiya, *Advanced Econometrics*, chapter 10, and in King, *Unifying Political Methodology*, chapter 9. The program itself is fairly straight forward— no effort is made to make it as sophisticated as it might be. It simply does the job.

The code comes in three main parts. First is a model specification section, that also sets up file names and variable names. Second is the specification of the log-likelihood function, in `proc llf`. Third is the actual call to the ML estimation routine, `Maxlik`.

While the program may look forbidding, this is more appearance than reality. Once you know how to write equations in *Gauss'* syntax, the rest is pretty simple.

Feel free to contribute short *Gauss* programs for future columns.

```
/*
**   A Type 2 Tobit Model
**   Stochastic Censoring Tobit
**   Amemiya, Advanced Econometrics, pp. 385-389.
**
**   by Charles H. Franklin
**   Washington University
**   March, 1991
**
**   Note:
**      y = x1*b1 + e1 is the OUTCOME equation
**      y2* =x2*b2 + e2 is the SELECTION equation.
**      Amemiya's notation reverses these.
*/
library maxlik.132, gauss.132;
#include maxlik.ext;
maxset;

/********* Model Specification Section *********/

__title = "Stochastic Censoring Tobit Model";
output file = tobit2.out reset;
datafile = "tobit2";
nx1 = 2;     @ number x1, NOT counting intercept @
nx2 = 2;     @ number x2, NOT counting intercept @
let dep = y;
let ind =x1 x2 x3 x4;
vars=dep|ind;
_mlparnm = "int1"|"x1"|"x2"|
           "int2"|"x3"|"x4"|"s1"|"s12";

       @ Appropriate starting values here@
let theta0 = 2 2 2 1 1 1 1 .1;

/****** Likelihood Specification Section ******/

proc llf(b,z);
    local y,x1,x2,beta1,beta2,s1,s12,u,a1,a2,a3;
    s1 = b[rows(b)-1];
    s12 = b[rows(b)];
    beta1 = b[1:(nx1+1)];
    beta2 = b[(nx1+2):(nx1+2+nx2)];

    x1 = ones(rows(z),1)~z[.,2:(nx1+1)];
    x2 = ones(rows(z),1)~z[.,(nx1+2):(nx1+nx2+1)];
    y  = z[.,1];
    u = y .== 0;

    if s1 <= 1e-4;
      retp(error(0));
    endif;

    a1 = x2*beta2;
    a2 = (a1 + (s12./(s1.*s1)).*(y-x1*beta1))./
         (sqrt(1-(s12.*s12)./(s1.*s1)));
    a3 = (y-x1*beta1)./s1;

    retp(u.*ln(cdfnc(a1)) + (1-u).*ln(cdfn(a2).*
          (1/s1).*pdfn(a3)));
endp;

/********* Maximization Routine Section *********/

_mlcovp=3; @ Compute White's H-C standard errors @

{theta,f0,g,h,rc} =
       maxprt(maxlik(datafile,vars,&llf,theta0));
output off;
```

The 1991 ICPSR Summer Program

Henry Heitowit
ICPSR

Members of the Political Methodology Section of the American Political Science Association may be interested in some of the recent innovations and additions to the courses offered by the Inter-university Consortium for Political and Social Research (ICPSR) Summer Program in Quantitative Methods of Social Research

A new course initiated last year is *Likelihood Models and Statistical Inference* which emphasizes maximum likelihood estimation and is based around Gary King's new book *Unifying Political Methodology*. The course will be team taught by Gary King, Victoria Gerus (both of Harvard University), Nancy Burns (University of Michigan) and others.

Another new course is a lecture series *Dynamic and Longitudinal Analysis*. This course will devote one week each to the following topics: Panel Analysis (Greg Markus, University of Michigan), Pooled-Time Series (Markus), Event History Analysis (Charles Denk, Sociology, University of Virginia) and Vector Autoregression (John Williams, Indiana University).

For the last several years we have offered three courses in the general area of "mathematical modeling:" *Game Theory* (Jim Morrow, Hoover Institution), *Rational Choice* (Jack Knight, Washington University), and *Formal/Dynamic Models of Social Systems* (Courtney Brown, Emory University).

Other recent additions to the Program include one-week (5-day), intensive courses on *Network Analysis* (Stanley Wasserman, Psychology, University of Illinois), *Regression Diagnostics* (John Fox, Sociology, York University), *Logit and Log-Linear Models* (Mike Berbaum, Psychology, University of Alabama), and *General Structural Equation Models— Introduction and Advanced Topics* (Ken Bollen, Sociology, University of North Carolina).

In addition to the above, there are the traditional 4-week Program offerings in such areas as Causal Analysis, "LISREL" Models, Time Series, Categorical Analysis, and Advanced ANOVA models.

The advanced topics (guest) lecture series this year will include presentations on "Chaos" and Non-linear Dynamics, Resampling Techniques: Jackknife and Bootstrap (Bob Stine, Statistics, Wharton School), Smoothing Functions (Werner Stuetzle, Statistics, University of Washington), and Graphical Presentation and Analysis (Bill Cleveland, AT&T Bell Labs).

The ICPSR Summer Program dates are July 1-July 26 for the first session, and July 24-August 23 for the second session.

The Summer Program curriculum is guided by an Advisory Committee composed of Chris Achen, Greg Markus, Jim Stimson, Ken Bollen, John Fox, and Cliff Clogg.

Individuals interested in receiving the Program brochure and an application should contact: ICPSR Summer Program, P.O. Box 1248, Ann Arbor, MI 48106 (313/764-8392).

I would like to take this opportunity to publicly thank Chris Achen for serving as academic/curriculum advisor to the ICPSR Summer Program for the last five years. During his tenure in this position the Program has experienced dramatic growth in the number and diversity participants. This was due in no small way to the contribution Chris made to our curriculum. His assistance, wisdom and wit are greatly appreciated.

Event History Summer Course

Paul D. Allison
University of Pennsylvania

A five-day course on event history analysis will be offered July 22-26 in Philadelphia. The instructor is Paul D. Allison, Professor of Sociology at the University of Pennsylvania. He is the author of the Sage monograph *Event History Analysis*, and has conducted this course for the past five summers.

The course will emphasize models for longitudinal event data in which the rate of event occurrence is a log-linear function of a set of explanatory variables. Topics include censoring, accelerated failure time models, proportional hazards models, partial likelihood, time-dependent covariates, competing risks, repeated events and discrete time methods. Participants will get hands-on experience with IBM-AT's.

The fee of $700 covers all course materials, but does not include lodging or meals. For further information contact Paul D. Allison, 3718 Locust Walk, Philadelphia, PA 19104-6299, 215-898-6717, ALLISON@PENNDRLS.BITNET.

Preliminary Program— 1991 Summer Methodology Meetings

The Eighth Annual Political Methodology Conference will be held Wednesday July 17 through Sunday July 21, 1991, at Duke University. The following is a preliminary program which is subject to change. Scholars not on the program are welcome to attend the sessions, though no support for meals or lodging can be provided.

The Ninth meeting is tentatively scheduled for Harvard University, in mid-July of 1992. Watch *TPM* for a call for papers in the fall.

The following are not necessarily in the order of the final program.

Session 0

"Spectral Methods for Time Series Analysis: A Pedagogical Session" Mel Hinich, University of Texas, Austin

Session 1

"Models of Voter Uncertainty", Charles H. Franklin, Washington University, St. Louis

"Messages Received: The Political Impact of Television News", Larry M. Bartels, University of Rochester

Discussants: Stanley Feldman, SUNY-Stony Brook; Jonathan Nagler, Texas A&M University

Session 2

"God's Model ...", John Freeman, University of Minnesota

Discussant: Michael MacKuen, University of Missouri-St.Louis

"Practical Reasoning: An Alternative Unifying Theme for Political Methodology", Hayward Alker, MIT

Discussants: Gary King, Harvard University; Phil Schrodt, University of Kansas

Session 3

"Time-varying Long Cycles in International Relations", John T. Williams, Indiana University

"Rational Expectations, Partisan Politics and Aggregate Economic Performance", Timothy W. Amato, University of Washington

Discussants: Nathaniel Beck, University of California, San Diego; Jim Granato, Michigan State University

Session 4

"MDS as a Statistical Method: Theory and Examples", Henry E. Brady, University of California, Berkeley

"Analyzing the Effects of U.S. Local Government Fiscal Activity", Walter R. Mebane, Jr., Cornell University

Discussants: John Jackson, University of Michigan; Simon Jackman, University of Rochester

Session 5

"Estimating the Effects of Campaign Spending in House Elections Using Semi-Parametric Methods", Donald Green, Yale University

"Bootstrapping: Computationally Intensive, Nonparametric Statistical Inference", Christopher Z. Mooney and Robert Duval, West Virginia University

Discussant: Douglas Rivers, Stanford University

Session 6

"Hierarchical Probit Models of Legislative Representation", Elisabeth Gerber, California Institute of Technology

"Bayesian Voter Models and Ecological Inference", Christopher H. Achen, University of Michigan

Discussants: William T. Bianco, Duke University; Arthur W. Lupia, University of California, San Diego

Political Analysis Available at 25% Discount

Those still looking for a good reason to join the Political Methodology Section of APSA may wish to look no further. The University of Michigan Press is offering section members a 25% discount on purchases of *Political Analysis*, the annual official publication of the Methodology Section. The discount alone ($10.62) more than pays for the section dues ($8.00). Not to mention you get a free subscription to *The Political Methodologist*. What more could one want from life? C'mon! Send in those dues and read all of *Political Analysis* for 3/4 the price.

The Political Methodologist
Department of Political Science
Washington University
One Brookings Drive
St. Louis, MO 63130

The Political Methodologist is the newsletter of the Political Methodology Section of the American Political Science Association. We gratefully acknowledge the support of Washington University in helping to defray the editorial and production costs of the newsletter.

Subscriptions to *TPM* are free to members of the APSA's Methodology Section. Please contact APSA to join the section. Dues are $8.00 per year.

Submissions to *TPM* are welcome. Articles should be sent to the editor, Charles H. Franklin, Department of Political Science, Washington University, One Brookings Drive, St. Louis, MO 63130. If at all possible, please send articles via Bitnet or Internet to the editor at C38871CF@WUVMD. Submissions may also be made on diskette as plain ascii files. TEX format files are especially encouraged, though we can read most word processor files. The deadline for submissions for the next issue is October 1, 1991.