The Political Methodologist

Newsletter of the Political Methodology Section
American Political Science Association
Volume 8, Number 2, Spring, 1998

Editor: Jonathan Nagler, University of California, Riverside

Contents

Notes From the Editor

Jonathan Nagler
University of California, Riverside
nagler@wizard.ucr.edu

This is my last issue as editor of *TPM*. It has an article by Frank Baumgartner, Bryan D. Jones, and Michael C. MacLeod on how to conduct a large data gathering project. This is one of the things that does not get taught, but that alot of us want or need to know. There is also an article by Sussane Lohmann on some practical, and theoretical, issues on how to conduct experiments. I'd like to thank everyone who has contributed to *TPM* over the last several issues. No doubt the new editors, John Londregan (jbl@Finer.sscnet.ucla.edu) and Mohan Penubarti (mohan@ucla.edu), will maintain *TPM*'s high standards. Finally, there is still space available for the short courses listed at the end of this issue!

Lessons from the Trenches: Ensuring Quality, Reliability, and Usability in the Creation of a New Data Source[1]

Frank R. Baumgartner
Texas A&M University
FrankB@polisci.tamu.edu

Bryan D. Jones
University of Washington
bdjones@u.washington.edu

Michael C. MacLeod
Texas A&M University
mac@polisci.tamu.edu

Introduction

For the past five years, we have been involved in the construction of a series of extremely large and, we hope, high quality data sets that should be of interest to a wide range of users in political science and public policy.* As we will discuss in detail below, these are very large datasets, designed to cover the universe of public actions in several areas: hearings, stories about congressional activities in the *Congressional Quarterly Almanac*, public laws, federal spending by function, and a sample of stories on all topics in the *New York Times* Index. Each covers the entire post-war period, and each is quite substantial by itself (our congressional hearings dataset, for example, includes 70,000 records, each of which includes 20 variables, and measures about 24 megs in a compressed Microsoft Access database file). In the process of trudging through these trenches over such a long period, we have confronted a number of problems common to all those who work to create systematic and reliable indicators from publicly available data sources. In what follows, we give some background into what led us into this

[1]**Research was supported by National Science Foundation grant # SBR-9320922 and by the Department of Political Science at Texas A&M University.**

project, what we were looking for, and we proceed with a number of lessons. A few of these lessons are mentioned in research design and methodology texts of which we are aware, but most seem to be the type of thing that may distinguish those who have been around the block a few times from those who have not.

We begin with a review of why we got into the project, a combination of our own interests and our desire to review publicly available records and to create a series of datasets that would be of interest to and be used by a wide range of users. We describe the nature, scope, and content of the five datasets, before moving on to a series of lessons having to do with how one ensures high quality in such a large project. First, we focus on distinguishing between those variables that are problematic to code from those that are straightforward. This determines the degree of oversight and reliability checking that must be invested. Second, we note the importance of admitting defeat: If it becomes clear that a variable has serious problems of reliability, and if this cannot be resolved with available resources, there is little merit in publishing the variable. We discuss how we came to admit defeat in a few areas of our project. We then consider some important software and data management lessons that we learned, in particular the importance of flexible data entry forms that allow coders to note any problems or ambiguities that they encounter as they run across them. This allows supervisors to identify problem cases and to resolve the difficulties systematically. It also allows successive waves of oversight to take place "upstream" from the original data entry, a process that ensures that the most problematic variables are scoured many times for potential reliability problems while others can be reviewed fewer times. We discuss the importance of oversight and creating an atmosphere of problem-solving among the project staff; of monitoring progress and difficulties as the project goes on so that one can ensure that resources are devoted to those areas that need them; how one can build in flexibility to the coding schemes so that the datasets will be useful to the broadest possible set of users; and, finally, how important it is to structure work on such a large project so that each staff member gains valuable lessons from the work.

Background and Goals

In 1989 two of us began collecting data in earnest for what became *Agendas and Instability in American Politics*. We noted that one could collect large amounts of data over long periods of time by using historical records such as congressional abstracts and media indices. Using simple spreadsheets, we had coders read hearings, for example, typing into the spreadsheet simple identifying material (date, CIS source code, committee name, title of hearing) as well as a short description of the hearing. From this initial data entry, which could be done by students who did not need extensive training, we could have more highly trained coders note whether the general topic of the hearing was favorable or negative towards the industry in question, or whether this tone could not be discerned. The coding of these tones proved to be surprisingly reliable and simple to perform. Further, the technique of having "front-line" coders in the library with portable computers entering data from these sources meant that their work could be supervised from a distance and that subsequent coding could be done from the spreadsheets and the short descriptions of the hearings that they typed in, rather than from the hearings themselves. This labor saving procedure meant that if we trained a small number of people to understand our coding processes in great detail, we could have a larger number of coders who would not need to be as extensively trained, and whose work would not involve making any sort of complicated coding decisions.

Using very simple spreadsheets, we coded thousands of congressional hearings, *Readers' Guide* entries, and *New York Times Index* summaries. By collecting a small amount of information from each of these sources, we were able to cover a long historical period, noting sometimes some surprisingly quick changes in tone and coverage of media and congressional attention. The key methodological innovation was the application of very simple coding rules combined with a coding mechanism that allowed many data entry personnel to be backed up and supervised by a much smaller number of more highly trained coders. This allowed reliability to be much higher than we could have achieved if each coder had worked independently.

At the completion of this project, we recognized that these methods could potentially be further refined and applied to a much larger research project: To code and to make available the entire public record from a number of sources, as described below. We had our own theoretical reasons for investigating government actions on a global scale, but we also expected that a broader group of scholars would find the resulting data sets to be of use. We proposed to the National Science Foundation the construction of a series of five linked data sets, each covering the entire post-war period. Now that the project is complete, we offer some lessons here.

The Datasets

The Policy Agendas Project involves the creation of five related databases as described in Table 1. The datasets include all congressional hearings, all public laws, all CQ stories, the complete federal budget by subfunction, and a sample of the *New York Times Index* for the post-war period. (We sample the *Times* because the census of all stories in the period would number in the millions.) The congressional

hearings, CQ stories, and federal budget datasets are currently available. For these datasets and a complete description of the variables included in each, see our project web site at http://weber.u.washington.edu/ ampol/agendasproject.html. The other datasets will be available by July 1998.

Table 1. Summary of Data Sets

Data Set/ Period Covered	Source/ Unit of Analysis	No. of Cases/ No. of Vars.
Congressional Hearings/ 1946-1994	CIS Abstracts/ Hearing	74,000/ 20
Public Laws/ 1948-1994	CQ *Almanac*, Appendix listing all public laws/ Public Law	16,300/ 17
Congressional Quarterly *Almanac*/ 1948-1994	CQ *Almanac*/ Story	12,600/ 37
Budget/ 1947-1994	*Budget of the United States* (annual)/ OMB subfunction	51/ 75
New York Times Index/ 1947-1994	*New York Times Index* (annual)/ Story abstracts	45,000/ 20

Four of the datasets have a common topic code that can be used to assess institutional attention to issues across the different archival source (links between the budget and the other data sets are incomplete because of the OMB classification schemes, as noted below). This is perhaps one of the most innovative and useful features of this project, so we devote considerable attention to it. Table 2 provides a list of the major topics used in these datasets.

Within each major topic, there are number of subtopic divisions. Thus analysis can be conducted on both a relatively high level of aggregation and a lower, more specific, level of aggregation. The topic code consists of four digits: the first two digits refer to the major topic and the last two refer to the subtopic. Table 3 gives an example of the subtopics for one of our major topic areas, heath care. Across all of our topic areas, there are about 225 subtopic codes.

Table 2. Major Topic Codes

1	Macroeconomics
2	Civil Rights, Minority Issues, and Civil Liberties
3	Health
4	Agriculture
5	Labor, Employment, and Immigration
6	Education
7	Environment
8	Energy
10	Transportation
12	Law, Crime, and Family Issues
13	Social Welfare
14	Community Development and Housing Issues
15	Banking, Finance, and Domestic Commerce
16	Defense
17	Space, Science, Technology, and Communications
18	Foreign Trade
19	International Affairs and Foreign Aid
20	Government Operations
21	Public Lands and Water Management
23*	Culture, Entertainment, Miscellaneous
24*	State and Local Government Activities

*Note: Topics 23 and 24 are used only in the *New York Times* dataset.

Table 3. Health Subtopics

300	General (includes combinations of multiple subtopics)
301	Health Care Reform, Health Care Costs, Insurance Costs and Availability
303	Medicare and Medicaid
306	Regulation of Prescription Drugs, Medical Devices, and Medical Procedures
307	Health Facilities Construction and Regulation, Public Health Service Issues
309	Mental Illness and Mental Retardation
310	Medical Fraud, Malpractice, and Physician Licensing Requirements
311	Elderly Health Issues
312	Infants, Children, and Immunization
313	Health Manpower Needs and Training Programs
315	Military Health Care
332	Alcohol Abuse and Treatment
333	Tobacco Abuse, Treatment, and Education
334	Illegal Drug Abuse, Treatment, and Education
349	Specific Diseases
398	Research and Development
399	Other

All of the datasets are coded consistently by topic or by subtopic, with two exceptions: The budget data set is coded by 72 OMB subfunctions, and these do not match our topic coding system completely; and the New York Times data are coded only by major topic, not by subtopic (see the section below entitled "Admitting Defeat").

More important than our coding system to many users may be the fact that our procedures for coding relied on the inclusion of a textual summary variable in which we typed in a short description of each story, hearing, law, or abstract. We used these summaries to code each item according to our scheme, but other users might well prefer to recode our datasets according to their own needs. One of the main reasons we included the textual summary, as mentioned above, was to allow us to separate the data entry process from the coding process so that we could have a smaller number of highly trained coders to memorize our complicated coding scheme. However for the user community this also provides some dramatic benefits in terms of flexibility. Table 4 provides examples of records for a health subtopic from each of the datasets.

Our datasets combine a set of complicated topic codes with a textual summary that should make them usable to a broader audience. Further, this additional flexibility actually facilitated, rather than hindered, the process of coding.

A Hierarchy of Hassle

Our main concern was to insure that whatever information we included in our data sets was accurate. Over the five years we worked on the project we have supervised dozens of students, both undergraduates and graduates. Some worked for us for years; others only for a semester or two. Some became intimately involved in the project, coauthoring articles using the data, writing dissertations related to the project, and knowing parts of it in great detail. Others became only marginally involved by working only on the front-line data entry. Just as we had a range of students with differing levels of involvement, so we also found that some parts of the project were trivial to oversee but that others required extensive training and checking for accuracy and reliability. We have come to think of this as the "hierarchy of hassle." Some variables were no hassle; others required a lot of work.

At the top of our hierarchy of hassles was the topic coding system used to link the datasets. Our initial concern involved developing a topic coding system that was flexible enough to allow meaningful comparisons across the archival sources. We attempted to develop a system of assigning topic and subtopic codes that avoids the assignment of thousands of distinct codes that would make analysis impossible, at the same time as we have tried to avoid the creation of so few topic codes that they might be home to too diverse a selection of entries. The trade-off, then, is between heterogeneity of topic categories on the one hand, and the multiplication of thousands of code categories, each of which might have so few entries that no analysis at that level of specificity would be possible. The other requirement, of course, is that the codes be clear enough that reliability be maintained.

Table 4. Selected Textual Summaries from Four Data Sets

Topic Code	Entry Summary
	Hearings
301	Federal health care spending
301	Health care reform and the role of medical technologies
301	Health maintenance organizations and hospitals providing managed health care
301	Health care access problems of disadvantaged and minority persons
301	Hospital financial practices and issues
	CQ Stories
301	Minority health: a non-controversial draft bill to authorize at least $144 million in fiscal 1994 to improve the health of minorities
301	Alternative health-care proposals: alternative plans made by Congress as opposed to the Clinton plan
301	Health care debate takes off: Congress gets up to speed on the complex economics and policies driving the US health care system
301	Health care program with included tax increase on the wealthy
301	Health care reform bill to impose national limits on health spending and expanded access to health insurance for pregnant women, children and those who worked for small businesses
	Public Laws
301	Amend the Public Health Service Act to provide an improvement in the health of members of minority groups
301	Provide federal assistance in establishing and expanding health maintenance organizations
301	Revise and extend the program for the establishment and expansion of the health maintenance organizations.
301	Enact the health maintenance organization amendments of 1978
	New York Times Index
3	Pres. Clinton's plan to save $35 billion from Medicare over next four years
3	Column article on both governmental and employers' long-term care policies and state intervention
3	Cost of health services should be distributed uniformly in all the states by financing it nationally
3	Hillary Clinton will appear before five committees of congress during hearings on Admin's health care plan
3	Letter from Western Pennsylvania Blue Cross executive officer explains how Penn. keeps percentage of people without health insurance under 10 percent

In coding over 70,000 congressional hearings by topic area, we developed a system that allowed us to code each item into one of 19 major topic codes and into 225 minor subtopic codes, as described in Tables 2 and 3 above. This

project started with congressional hearings, so we first developed our topic coding system on these data. Due to the importance of the topic code and the potential for intercoder reliability problems because of the large number of student workers and the complexity of the coding system, we decided that two students would be responsible for all topic assignments. To do this, we had the original coders include a short description of each hearing in the dataset. Then our two "expert topic coders" assigned topics based on these descriptions. This greatly reduced the potential for error and allowed us to recheck the codes later.

Once a few years of hearings were complete our two experts began separately assigning topic codes to the same entries. This served two purposes: (1) we assessed intercoder reliability; and, (2) we substantially revised and refined our codebook. Each week, the two experts would code several hundred hearings and then meet with Baumgartner or Jones to resolve differences and amend the codebook. We did not allow the coders to work on their own until reliability scores were at or above 90% for the major topics and 70% for the subtopics. Over a period of two months we met these goals. Over the first two years, our reliability scores with new coders rose from 85 to 95% on the major topics and from about 65 to 85 or 90% on the subtopics, based on periodic tests of 100 cases.

Our preliminary checks on reliability were used only as a means of assessing when a coder could begin to work in earnest. Our data cleaning procedures insured that each variable toward the peak of the hierarchy of hassle would be reviewed many times before the data sets were released; those at the low end (for example, those items that were simply recorded verbatim, such as dates) were reviewed only twice for the most part. In the case of the topic codes, periodic reliability checks combined with weekly meetings led to continual revisions of the codebook. After 10,000 hearings were complete, they were sorted by topic and the reviewers checked the entries to insure that they were correct and to make updates consistent with revisions. Here, problematic subtopics were excluded and several of the major topics were combined (e.g., Law and Crime was combined with Family Issues; and, Illegal Narcotics was divided into Health and Law and Crime).

At this point several problem areas remained. Certain groups often mentioned in congressional hearings, such as the elderly and veterans, were very difficult to code because of overlap between social classification and substantive topic (here we wished that we had allowed for two topic codes-see discussion below). We decided to include separate subtopics for each of these groups in different substantive areas. For example, there are subtopic codes for the elderly in civil rights, health, social welfare, and housing. Veterans' issues were divided between health, housing, and defense. This allows scholars studying the elderly or veterans issues to simply combine each of these subtopics for analysis. In sum, we faced a dilemma about priorities: Is a hearing about health care for the elderly properly coded as health care, or as elderly issues? Is veterans' education benefits a military issue or an education issue? We resolved these issues by creating a subtopic code for each such potentially confusing category. We code elderly health care issues under the major topic of health care, but it is trivial for another analyst to recombine our subtopic codes to conform to whatever logic they might like. In this way, our four-digit subtopic coding serves our purposes, but we think that the specificity of the 225 subtopics is great enough to allow subsequent users to recombine them in almost any way. (Further, because we include a textual summary of each entry in all of our data sets, subsequent users can create even finer classification schemes than we use: they merely start out with our subtopics that include the narrow range of issues they are interested in, then read through our summaries to extract and recode any particular entries of interest.)

The other major problem involved dealing with about 5-10% of the cases that seemed to include two distinct major topics. In our coding for *Agendas and Instability* we allowed for up to five topic codes for each hearing. We found that the vast majority of hearings could be identified with only a single topic identifier (over 90 percent), and the results of any analysis conducted on only the first topic code differed only slightly from those that we conducted on all five possible topic entries. Based on this experience, we streamlined the data collection in this project to allow for only a single topic descriptor for each hearing. Here we simply imposed a rule stating that these cases should be coded under the topic code that came first in our topic coding system. We also found that many of the hearings discuss several subtopics within a major topic area, but this problem was easily handled by noting a special subtopic code, indicating "general, or several subtopics discussed."

Still, scholars might argue that we have lost a substantial amount of information by forcing some records into a single topic. We got into a rather heated debate over this problem while presenting a conference paper on the congressional hearings data. One scholar rightfully noted that some topics such as environment, energy, and health care are more likely to include multiple dimensions of debate, and thus might require more than one topic code. Again, our solution to this potential problem was to include a short written description of each record in each dataset next to the topic code. Thus users not only have the option of sorting the dataset by subtopic to verify the accuracy of our topic coding, but also they can recode the entries using multiple topic codes for each record. We felt that this approach allowed us to accomplish our goals, while at the same time

preserving as much information as possible for use by others. This flexibility is important in maintaining the historical value of the dataset and in insuring that it may be useful to different end-users.

After all hearings were coded, and after continual weekly meetings, the two senior project leaders and the two expert coders reviewed all 70,000 topic entries. We sorted the hearings by subtopic (so that all the hearings coded as being in the same subtopic category would be listed in order) and we painstakingly compared the short description of each entry to the topic code. In this way we checked for errors and made some last minute changes to ensure that each subtopic was relatively homogeneous. Because we had the summary descriptions in the data sets as well as the numeric codes, we could use the data base software to print out the records in sorted order and to revise our topic coding when we saw discrepancies. In this way, we revised our topic coding continually during the first months of the project, then checked it through completely for consistency at the end. Further, we made sure to revise the printed codebook so that it reflected the changes as we made them. Needless to say, this took over 5 months and all required bifocals after the experience. (When we repeated these procedures for the data sets to come on line later, Jones had moved to another university, so Baumgartner and MacLeod had the exclusive privilege of reading each public law entry, each story in the CQ *Almanac*, and each *New York Times* story in our sample.)

We used the same topic coding system for public laws and CQ stories abstracts as we developed for the congressional hearings. Inevitably, because public laws tend to be more general than hearings, we made more use of the "general" subtopic within each major topic area in these data sets than we did in our data set on congressional hearings. In terms of the 19 major areas of public policy that we code, our data on public laws and CQ stories are consistently coded with inter-coder reliability scores of 96%; for the subtopic codes, reliabilities are above 90%. After coding over 28,000 cases from 46 annual editions of the CQ Almanacs (and 70,000 hearings in a previous project), we are confident that our topic and subtopic codes produce useful and homogeneous groupings.

Another variable towards the peak of our hierarchy of hassle was the committee and subcommittee code. There are no theoretical issues here, only a logistical nightmare stemming from the fact that Congress often reorganizes the subcommittee structure. In order to trace the flow of issues through committees and subcommittees over time, we had to account for the periodic reorganizations and continual subcommittee name changes. Unlike congressional committees, there is no source that documents subcommittee changes over time. *Congressional Quarterly Almanacs* and *Congressional Staff Directories* were used to track name changes and dates of activity for subcommittees. These materials provide yearly lists of subcommittee names and their members. Subcommittee name changes were matched through time by comparing the membership lists of each subcommittee from year to year, a rather formidable task considering that our codebook contains nearly 1000 subcommittees. In some cases it was clear from the membership lists that an existing subcommittee had simply changed its name, in other cases it was unclear. If we could not determine that a subcommittee was new or an existing subcommittee that had been renamed, we simply treated the subcommittee as new by assigning it a new code.

Despite our efforts, there were inevitable discrepancies between our codebook and the subcommittee names listed in the CIS hearing abstracts. If our coders could not find the subcommittee name listed in the abstract in our codebook, they were instructed to assign a new code to it and then to write it in the codebook. Codebooks were turned into the hearings project manager at the end of each completed volume. The project manager then updated the codebooks and redistributed them. At the completion of the project, a massive set of crosstabs (subcommittees by year) were run to check for errors. Here we looked for subcommittee numbers that did not exist in our codebook and for subcommittees that held hearings more than several years apart. Using this procedure we identified about 6000 hearings that needed to be checked. We assigned a single person to go back to the original abstracts and verify that the subcommittee number matched the name in the abstract. This procedure resulted in some minor adjustments to the data and codebook. In hindsight, we think that it might have been easier to have the coder type in the name of the subcommittee rather than assigning a new code. The project manager would then review each year and assign the new codes.

While the subcommittees and the subtopics were very time consuming and required extensive checking for accuracy, the vast bulk of the data collected in our datasets were extremely straightforward to code. Our web site includes a list of all the variables for each of our data sets. Besides the textual summaries that should be of use to others, we also were careful to include complete identification for each entry so that any user wishing to go into more detail than we did from the original data sources can located the records. Further, we note a number of yes/no variables that were trivial to collect once our coders had read the story, description, or abstract. Examples of this are whether a *New York Times* story mentions anything about: the President; Congress; state and local government actions; courts; elections and campaigns; etc. Our hierarchy of hassle had a few complicated items towards the top, requiring extensive project resources to get right, but it also include a broad

base of information for which we had very little difficulty in maintaining accuracy.

Admitting Defeat

From the outset, we decided to exclude variables unless we could guarantee a high level of reliability. Rather than include in our data sets any variables that we were not confident of, we found in a few cases that we simply had to admit defeat. Given Jones' background and interest in urban studies we originally included a variable in the hearings dataset that was designed to let us identify all hearings that involved urban issues. At first, we developed a list of keywords (e.g., community development, mass transit, homelessness, etc.) that coders could use to identify these hearings in order to reduce reliability problems. While this worked well, we soon discovered that the list led to massive reliability errors that made us suspect the validity of the measure. After months of project meetings in which we asked each coder whether they would code certain cases as mentioning urban issues, we finally decided that there was indeed no systematic and reliable way to make this distinction, and we reluctantly deleted the variable. It just didn't work.

A similar problem arose with a variable in the *New York Times* dataset designed to capture articles that dealt with a social category. We originally wanted to be able to analyze stories by social groupings, but it was often unclear from the stories if the group mentioned belonged in this category. It was clear enough if a story mentioned an organized interest group (a variable that we do code), but we also wanted to note if the story mentioned any social or demographic group such as the elderly, ethnic groups, or others. To our dismay, we found that some coders understood this differently than others and we were unable to write a simple rule that would have sufficient validity. We dropped the variable.

Having started with the collection of congressional hearings, then proceeding when that was nearly complete to develop the CQ Stories, budget, and Public Laws data sets, we came last to the *New York Times*. Where all the other data sets concerned government action, of course media coverage is much broader. Further, the subtopics discussed even within the same major topics tend to be quite disparate. Rather than stretch the definitions of our subtopics so that they might mean different things depending on the data source, and rather than multiplying our subtopics to a dizzying number, we simply decided not to code the media data by subtopic at all.

Finally, we had hoped to link the budget data to the other datasets using the same topic and subtopic codes. Unfortunately, the Office of Management and Budget does not think the same way we do. There are clearly a number of cases (representing over half of the dollars in the federal budget) where the OMB classifications correspond very well with our major topics. However, in the areas where they do not, we could not break apart their spending categories to correspond perfectly with our subtopic codes. After spending two years on the creation of this data set alone, we finally decided to declare a partial victory and to move on to other tasks. No amount of effort would have solved this problem completely, but our data sets still correspond in enough cases to be useful.

Software and Data Management

As one can imagine when looking at the scope of our project, one of our most pressing and constant concerns was data management: insuring that we lost no data, that we could distinguish between preliminary datasets that had not yet been checked for errors and those that had, and the like. Our data management problems grew as the project progressed. Initially we coded congressional hearings in Microsoft Excel, but we quickly exceeded its sixteen-thousand row capacity. We needed software with the ability to not only manage thousands of records but also one with a strong analytic capacity. Fortunately, we had a graduate student with a great deal of software knowledge, and he suggested that we try Microsoft Access. The learning curve was slow using this software, but it served our immediate data management needs because of its one-million-plus record capacity.

As our knowledge of this software progressed, it led to important improvements in subsequent projects. When we began coding CQ stories and public laws, we developed customized coding forms in Access that greatly reduce the potential for error and also increased data entry speed. These coding forms include the names of each variable and have a space for entry. Coders tab through the form as each variable is entered for a single record. Only one record at a time is displayed. Categorical variables can have the acceptable categories already available for the coder to choose. Dichotomous variables can be presented as check boxes with a default of whichever is the most common choice.

These forms reduce errors on several fronts. First, parameter values can be set for each variable. For example, the acceptable values for the variable "year" were 1947 to 1993. Parameter values can also be set to text or number. When an incorrect entry occurs, Access generates an error message that prompts the coder to fix the entry before he can proceed to the next variable. The time spent fixing errors was greatly reduced using this system. Second, from a data security standpoint, coders do not have access to the original file that contains the data. Coders open the file and the form appears showing the last record entered. Coders enter one record at time, and once entered, the data are in a table that only the project manager can access for revisions.

Another important feature of Access is that it allows us to link and analyze multiple datasets. By creating an Access database including tables with data for the hearings, CQ stories, public laws, and *New York Times* stories, we were able to run analysis on the number of records by topic for each data source. Access allows users to run analysis on two or more datasets as long as there is at least a single variable in common. Several graduate students on the project used some of the congressional hearing data that had been compiled and added witness information. This is difficult from a data management standpoint because it involves two different units of analysis (hearings and witnesses). This problem is easy to handle in Access by including a common variable in each dataset. In this case, the common variable was the CIS hearing number and this was used to link the datasets for analysis.

We instituted several programs for data backup. First, we worked with our department computer specialist to set up our own network server. We purchased a large hard disk that was devoted to project data. All faculty and students on the project had access to this server, but it was not available to others outside the project. MacLeod was primarily responsible for ensuring that the data were backed up on a weekly basis using tape back ups and eventually switching to a Zip and a Jazz drive as technology progressed. Jones forced MacLeod to sleep with tapes under his pillow at night for safekeeping. Data were also backed up on a hard drive of one of the project computers. Thus we always had two backups of the network drive, one in the department and one outside the department in case of some catastrophic event. With the release of the data, we have it simultaneously in many locations. Of course, we keep all the preliminary files used for data entry until after we are certain that they have been merged with subsequent files successfully.

What Problems Did You Encounter This Week?

One of our most important procedures was a mandatory weekly project meeting to assess progress and to discuss coding problems. Key to this was a software innovation in which we included, as the last variable for each entry, a 255-character "Notes" variable. Here, coders would write in any questions or problems and then we would either discuss and resolve these problems at the weekly meetings, or, after completion, a department head would sort the dataset by notes and try to fix the problems. Particularly problematic entries were finally solved by MacLeod, Baumgartner, or Jones. Notes were very detailed and, in most cases, questions could be resolved without referring back to the original entry. As coders submitted their work to their project manager, this person could quickly scan the notes to see where any problems occurred. Those that could not be quickly resolved formed the subject of the weekly meetings.

How Much Progress Did You Make This Week?

Baumgartner continually annoyed all involved in this project by constantly worrying that available resources would lead us to four or five incomplete data sets, of value to no one. Measuring the pace of work as it is being accomplished is no fun, but it can help ensure that resources are being allocated where they are needed. We focused on two elements here: Ensuring that no particular variables were requiring so much coding time that their inclusion was slowing down the project unacceptably (and, we noted, those that take a long time are usually the ones with the unacceptably low reliability scores); and adjusting the sampling rates for the one data set where we did a sample rather than a census. Keeping good track of progress was necessary also as we calculated how many staff we needed for initial data entry versus coding, merging, and other "back office" tasks. In any case, it was essential to our successful completion of five large data sets.

Encouraging Secondary Use of the Data

We worked carefully to ensure that our data sets would be of use to ourselves, but we also wanted to make them of use to the broadest possible set of secondary users. In addition to the summary text forms and the inclusion of all necessary identification material so that a user can look up each case in the original data source, as we discussed above, we have worked to ensure extensive documentation of all the data sources and lately we have been working on the development of an extensive and growing web site for the distribution of the data. As more and more scholars take advantage of the ability to collect large data resources, we each have an obligation and an interest in promoting the sharing and wide use of the data.

As we were dependent on published sources for our data, users need to know the flaws as well as the merits of these data sources. Since we were the ones who became most familiar with the primary source, we should explain to subsequent users the potential problems with their use. One thing we noted in general involved the diminishing quality of the archival data back through time. CIS hearing abstracts and the New York Times abstracts presented particular problems. From 1947 to 1969, both these sources include much less substantive information. Most of the entries only included keywords. This had an impact on several of the variables in our datasets. Variables designed to allow us to identify hearings that involved new agencies, new programs, and whether the administration was involved are coded as missing data from 1947 to 1969 because the published record was not sufficiently detailed.

The congressional hearings data set posed another problem having to do with "unpublished" hearings held in

executive session. These hearings were not published with the annual volumes, but CIS has published thousands of them in subsequent years, in special volumes. This process is not complete; several thousands additional hearings are likely to be published in the future as CIS continues its process of releasing this large backlog. We explain these details in our web site; the general point is that users should become familiar with the datasets they are analyzing even if they did not gather the data themselves. We suspect that many of the most innovative and effective uses of our data sets will come from those users who use some elements of our data to save themselves from replicating our efforts, but who add to these variables a range of new items. For example, several of our students have written papers or dissertations isolating a single policy area or issue for more intensive analysis. Starting with our data sets, they can select out all hearings, laws, etc. on whatever topic interests them, and they can go back to the source materials to gather more information, allowing them to construct a more complete legislative history or a more detailed chronology or adding an analysis of roll-call votes where they were able to isolate the key pieces of legislation from our larger data sets. In sum, these data should be of use to many, but they should be seen also as only part of a picture to which users will add their own touches.

Specialization, Motivation and Incentives

As the project progressed we jokingly noted how bureaucratized it had become. Baumgartner and Jones formed the top of the bureaucratic hierarchy, while senior graduate students were in charge of separate departments and each had their own staff. In this way, each department head became an expert in their respective areas: Each was responsible for progress reports, coder training, intercoder reliability, and data review and management. Within each department there were specialists as well. Some of the students just coded information, while others reviewed completed work for errors. These reviewers typically developed expertise in several variables. This hierarchy was also useful for data cleaning, checking for errors, and the like. Some variables required little review and so were checked by front line coders; those that required greater oversight were reviewed by variable experts and the department heads. Problems were referred to MacLeod and items that he could not solve were kicked up to Baumgartner and Jones. In this way, we ensured that important variables received several layers of review, while at the same time not burdening project heads with unnecessary work.

Assignments largely had to do with student interests and capabilities. Each project manager was assigned on the basis of their research interests, software knowledge, and their seniority. Efforts were made to match individual research interests with work on the project. Motivations and incentives for students varied greatly. Graduate students with related research interests were allowed to use the project data in their own conference papers, journal articles, and dissertations. Seven of the graduate students used project data in their dissertations supplemented with original data collection efforts of their own. Six graduate students were co-authors on conference papers, three graduate students were co-authors on journal articles, and one published a single author article using project data (several other manuscripts with graduate students are under currently under review or in preparation). For these students the experience was ideal, but for others the experience was less than exhilarating.

This project involved rather monotonous coding, and graduate students without specific research interests related to the project data found the work rather boring. Inevitable complaints arose about the boredom, but most of these students continued to work on the project because of the flexibility of the work schedule, the consistency of the tasks, and access to project computers. Typically, research assignments at Texas A&M involve variable duties, lots of contact with professors, and some of the research required is very complex; at times, it was beyond the capabilities of some of the students that we employed. Once our coders were trained, they were allowed to work independently in the library on their own schedule. We reserved a room in the library and equipped it with a desktop and a laptop computer for their use. They were only required to attend a weekly project meeting to turn in the data that they had collected and to discuss problems. In addition, graduate students had access to project computers, usually the most advanced computers in the department, for their personal work. Most of these graduate students stayed with the project for these reasons, but a few asked to be reassigned.

Undergraduate student workers were motivated by money, letters of recommendation (most who worked for us went on to graduate or professional school), and the opportunity to be involved in research; although we also found that they enjoyed the social nature of the working environment. Most of the students were friends and worked in our department lab. At times, they seemed to be spending as much time on gossip as they did on work. However, they turned out to be indispensable. One undergraduate worked so hard that we nicknamed him the "rate buster" because of his penchant for outworking and out performing most of the graduate students. In addition, hiring undergraduate workers was cost efficient though they typically did not stay with the project for as long as the graduate students. Most were hired in NSF's Research Opportunities for Undergraduates program. Finally, all of these students gained valuable research experience and developed proficiency using several types of software that they were able to put to good use in graduate school.

Finally, MacLeod notes that one of the most important benefits for graduate students was the exposure to a large research project. He and others on the project learned about grant writing, the development of codebooks and coding procedures, student oversight and management, data reliability and validity concerns, and data management and analysis. This experience was invaluable to the graduate students. It was used by some to apply for external dissertation grants, and led to several large original data collection efforts by student for their dissertations. Jim True used the budget data to study punctuated equilibria in the federal budget; Jeff Talbert used some of the congressional hearings and added mark-up data to study the health care policymaking; Doris McGonagle used congressional hearings and added interest group testimony to study the social construction of elderly policy; MacLeod is using some of the congressional hearings and coded interest groups witnesses to study rapid policy change (positive feedback) in telecommunications policy; Glen Krutz is using the statute data to studying the rise and impact of the use of omnibus legislation. Most of these students used Microsoft Access for data management, many developed their own coding forms, and some hired student workers to code additional archival data for their dissertations. Because of this experience, these students are producing higher quality dissertations with a greater potential for future publication, and they are better prepared to conduct their own large projects as junior professors.

Rational Choice in the Laboratory: A Survivor's Guide to Experimental Design

Susanne Lohmann
University of California, Los Angeles
lohmann@ucla.edu

When we submit papers to refereed journals, we do not generally report our research findings in the messy and convoluted way we arrived at them. Formal theorists typically experiment with many different combinations of assumptions before they settle down on a specific formal model; they do not always tell their readers which assumptions are critical. Empiricists typically run many more regressions than they document in their papers; their statistical tests do not usually take into account the degrees of freedom they de facto lose when they run any number of unsuccessful regressions until they find a specification that "works" (yields coefficients "significantly different from zero at standard levels of significance"). Experimentalists typically run pilot experiments to get rid of bugs in their experimental design; on occasion, failed experiments are ex post declared pilots. Computational social scientists run large numbers of simulations, varying the parameters of the model until one parameter constellation yields a pattern consistent with the empirical data they are trying to match; they do not always report on the robustness of their simulation results.

To some degree, these behaviors are unethical, to some degree, they are a normal and unavoidable part of doing science (Latour, 1987). As a practical matter, even a scientist who has the desire to be honest cannot document everything she did in excruciating detail: instead of writing a thirty-page research paper, she would end up writing a 10,000-page book that is unreadable and unpublishable (and even then she would fail to document literally everything).

Some of the information that allows scholars to understand, replicate, and build on each others' findings is communicated not as part of peer-reviewed publications but in verbal and electronic mail exchanges, in workshops, conferences, and internet discussion groups. This article is meant to serve a function somewhere between a fleeting and tentative verbal communication, on the one hand, and a definitive article published in a refereed journal, on the other.

I report how I came to test my theory of dynamic informational cascades experimentally; what went wrong; what went right; and why. I believe that the tools coming out of experimental economics have enormous potential for testing rational choice theories in political science. My goal is to give other formal theorists a sense of the practical problems and judgment calls involved. In the limited space available, I obviously cannot do a complete or systematic job. At the end of this article, I provide a list of useful background readings for those who want to know more.

About two years ago, I started out with the expectation that experiments would allow me to test my theory of dynamic informational cascades without the clutter of real-world confounding factors. I would gain control by designing the monetary incentive structure faced by the participants in my experiments so as to mirror the incentive structure faced by the agents in my formal model. The participants would respond to the monetary incentive structure and in effect play "my game." I would then check whether the participants' decision rules were consistent with the equilibrium prescriptions of my model.

In hindsight, my expectation was naive. The participants in experiments are real human beings who reason and act according to a logic that is at best loosely related to rational choice. It is impossible, I think, to design and run experiments and get decent results without paying attention to insights about human nature coming out of cognitive, social,

and evolutionary psychology. As a result of running experiments, I have come to the conclusion that rational choice is a very impoverished theory of human behavior. (This conclusion is blindingly obvious in hindsight, but I guess we all come indoctrinated out of graduate school.) I find myself embracing the idea that we need to move beyond the rational choice paradigm if we want to understand such diverse phenomena as collective action, electoral competition, legislative norms and organization, bureaucratic decision-making, international conflict and cooperation, and so on.

The next section summarizes my theory of dynamic informational cascades. The following section describes my experimental design. I then discuss some of the judgment calls I had to make in writing the experimental instructions, executing the experiment, and interpreting the experimental data. Finally, I discuss how we might move beyond the rational choice paradigm, using laboratory experiments and computational models to spell out systematically how and why human beings deviate from the normative prescriptions of game theory.

Theory of Dynamic Informational Cascades

My theory takes as a starting point a society facing a collective decision between the status quo and an alternative. The status quo is overturned if the number of people taking costly action exceeds a critical threshold; if too many people abstain, the status quo is maintained. Each individual has some information (good or bad) about the status quo, and she can choose to act or abstain in each period over the course of multiple periods. An action serves two functions: it may contribute towards overturning the status quo in a given period, or it may signal the actor's information and influence other people's decisions to act or abstain in future periods. People rationally take informational cues from the aggregate number of people taking action. A dynamic informational cascade arises when the incentives to act fluctuate across people and over time so that information is extracted from different subsets of the population over time. This process of information aggregation is shaped by incentive-compatibility constraints arising from conflicts of interest across people and the free-rider problem of collective action.

In earlier work, I use my theory to analyze the "Monday demonstrations" in Leipzig, East Germany, which played a critical role in the East German revolution of 1989 (Lohmann, 1994). I am now extending the theory to analyze demonstrations, emigration, public opinion polls, and elections in the German Democratic Republic, during the East German revolution, and in post-unification East Germany (Lohmann, Work in Progress).

Real-world data allow me to discriminate between my model and competing models to some degree, but a definitive test remains elusive. The data are driven by an uncountable number of factors that are neglected or kept constant in my model. This is, of course, a generic problem: by their very nature, formal models abstract from the real world. But my philosophy of "fat-free modeling" – the idea that a model should be stripped down to the point where a result of interest is generated by a set of minimally sparse assumptions – makes my work especially susceptible to the accusation that it is not empirically descriptive.

For example, in my model the cost of collective action is assumed to remain constant over the course of a cascade. This assumption plays an important role for the interpretation of my theoretical results. The herding effects in my model are endogenously driven by the revelation of information through collective action over time. If I allowed the cost of action to vary over time, then herding effects would arise for "safety in numbers" reasons, and it would be conceptually hard to disentangle the role of information aggregation and of safety in numbers. In practice, of course, the cost of collective action does change over time. It costs more to turn out in a demonstration when it rains than when the sun shines; it costs more to turn out if the incumbent regime is known to admire the Chinese government's handling of the Tienanmen Square protest than if the right to free assembly is well-established and respected. The mismatch between reality and the simplifying assumptions that enter my theory obviously creates problems when I take my theory to the data.

I turned to experiments with the expectation that experimental data would allow me to avoid real-world complications and make cleaner inferences. Even at the time, it was clear to me that an experimental test of a theory of collective action has natural limitations. As a practical matter, I cannot pack huge numbers of people into a laboratory. In the laboratory, my "small N" case involves three people, my "large N" case, fifteen people, with N standing for the number of people who have stakes in the collective enterprise. There is some question whether the experimental results for $N = 3$ and $N = 15$ carry over to thousands, tens of thousands, or even hundreds of thousands of people participating in mass action in the real world. Here, my expectation was that real-world and experimental evidence would complement each other, with real-world data suffering under a lack of internal validity (do the data allow us to disentangle whether the model is correct?), experimental data under a lack of external validity (do the inferences we draw from experimental data generalize to field data?). Together, I thought, these two types of evidence would allow me to make a strong case for (or, as the case may be, against) my theory of dynamic informational cascades.

Laboratory Experiments

Between February and May 1997, I ran nine experiments testing my theory of dynamic informational cascades. In four experiments (3/3/97, 3/10/97, 4/11/97, and 4/29/97), fifteen participants formed one fifteen-person group; in another four experiments (2/3/97, 2/24/97, 4/25/97, and 5/16/97), fifteen participants formed five three-person groups; in one in-class experiment (5/22/97), over one hundred participants formed one group. The participants consisted primarily of UCLA undergraduates.

Here is the description of a typical $N = 15$ experiment (4/29/97). The experiment runs over ten rounds, with each round consisting of five periods. Each round starts with the public toss of a (real) coin that selects one of two urns. One urn contains two green chips and one white chip (GGW), the other urn, two white chips and one green chip (WWG). The two urns are identical when viewed from the outside so that noone (including the experimenter) can see which urn was choosen. Each participant draws one chip from the chosen urn and privately observes the color of the chip. Then, in the first period, each participant hands in an anonymous ballot indicating whether she is casting a costly vote or abstaining. One (randomly selected) participant adds up the ballots and publicly announces the total number of votes. If at least eight participants cast costly votes, the status quo is overturned, and the round ends. Otherwise the round continues with the second period. The play of the game in the first period is repeated up to five times. If the status quo is not overturned by the end of the fifth period, then the round ends with the status quo being maintained. At the end of the round, the chosen urn is opened, and all participants can see whether it is the GGW urn or the WWG urn.

In each round, each participant can get a payoff ranging from zero dollars to \$1.50. If the status quo is maintained, then each participant receives one dollar if the GGW urn was chosen by the toss of the coin, zero dollars if the WWG urn was chosen. If the status quo is overturned, then each participant receives an individual-specific alternative payoff: one participant receives 25 cents, the second, 30 cents, the third, 35 cents, and so on, with the fifteenth participant receiving 95 cents. (The distribution of individual-specific alternative payoffs is publicly established at the beginning of the experiment, with each individual's stake remaining private, and it remains fixed throughout the experiment.) Moreover, at the beginning of each round, each participant receives five voting coupons worth 10 cents each; each time she votes, she has to give up one voting coupon. If she has voting coupons left over at the end of the round, she can exchange them for their cash value. Thus, by abstaining, each participant can make up to 50 cents per round.

The experimental design captures the essential features of my theory: (1) people have private information about a common value; (2) people have different stakes in the status quo relative to the alternative; (3) costly collective action is beset by a free rider problem; (4) individual actions and abstentions are anonymous, and only the aggregate number of actions and abstentions is public information.

For the special case of $N = 3$, my model makes point predictions about people's inclination to take costly action as a function of their private information, their individual-specific stakes in the status quo relative to the alternative, and the aggregate number of actions observed in previous periods (Lohmann, 1997a, Forthcoming). For the special case of $N = 15$, a closed-form solution is not available, but my theory places qualitative restrictions on the data (Lohmann, 1997b). My model, like any signaling model, is beset by the problem of multiple equilibria. It assumes, like any game-theoretic model that gives rise to multiple equilibria, that people magically coordinate on one equilibrium; it does not say how agents deal with or overcome the problem of strategic uncertainty (which equilibrium will be played?).

Judgment Calls

I now discuss some of the judgment calls I had to make in designing the experiment. The first step in running a game-theoretic experiment is to get cash. It is standard in experimental economics to attract student participants with the prospect of making money and to impose a monetary incentive structure that mirrors the incentive structure of the game-theoretic model. By way of contrast, psychologists often encourage (force?) students to participate in experiments by making participation a class requirement, and they expect the participants to follow their instructions in the absence of explicit monetary incentives. The approaches taken by economists and psychologies each have their upsides and downsides, and it exceeds the scope of this article to discuss the trade-offs in any detail. Let me spell out one example of the kinds of considerations "outside of rational choice theory" an experimenter needs to keep in mind.

Rational choice theorists tend to emphasize the importance of instrumental motives and formal incentives. Using money to motivate participants is considered uncontroversial. It is useful to keep in mind, however, that money may not be neutral because it interacts in subtle ways with noninstrumental motives and informal incentives. Suppose, for example, that some percentage of the faculty in your department attends departmental meetings on a regular basis. Now suppose one day the chair of your department decides to encourage attendance by paying \$100 per person and meeting. Attendance does indeed pick up. One year later,

the chair faces a budget crunch and is forced to end the payments. The question is what happens to attendance: will it stay at the new level, drop to the earlier level, or drop below the earlier level? I predict the latter. Before the chair reimbursed faculty for attending departmental meetings, faculty attended for mixed reasons, only some of them instrumental: to shape the future of the department, in response to group pressures (there are social rewards for attendance, punishments for non-attendance), and out of a sense of obligation. By providing yet another, monetary, reason for attending departmental meetings, the chair strengthened the incentives to attend, but he or she may have inadvertently weakened the group pressures enforcing attendance and undermined the sense of obligation to attend.

The rational choice paradigm is largely blind to the informal incentives shaping people's behavior, and it is outright hostile to the idea that people behave independently of the incentives operating on them. Good experimental design requires an open mind to subtle effects that arise outside of the rational choice paradigm rather than a dogmatic mind that seeks to ex post rationalize everything observable within the paradigm.

The second step in running experiments is to recruit participants. There is a vast literature on how to select or assign participants "randomly" so as to avoid selection biases. My sense is that experimental economists and experimental psychologists (and their political science equivalents) have different attitudes towards the question of random assignment. Experimental economists tend be rather cavalier, experimental psychologists prissy. The reason is, I think, that the rational choice paradigm implicitly conceives of people as being more or less identical in the sense that it expects two human beings faced by the same incentive structure to behave in the same way. There is nothing in conventional rational choice theory suggesting that an Asian American woman and a white male will behave differently when faced by the same monetary incentive structure. Because the a priori expectation that race, gender, and other individual characteristics do not matter is so strong, experimental economists rarely run surveys along with their experiments (their disdain for the ill-defined incentive structure created by the request to fill out a survey questionnaire also enters here). In contrast, psychologists have a view of human nature by which people's behavior is driven by internal factors, many of which are shaped by personal characteristics or individual-specific life experiences. Thus, psychologists often ask batteries of questions that to an economist appear totally irrelevant ("were your parents divorced?").

The third step in running experiments is to get a room and equipment. Here the big question is whether to use computers or to run the experiment by hand. Once again, there are upsides and downsides either way, and I don't have the space for a detailed discussion. Let me mention a few considerations I believe are especially important.

A social scientist who is running experiments for the first time should run them him- or herself by hand, that is, not delegate the work to a research assistant and not use computers. The details of what goes on during an experiment cannot be exhaustively communicated in word or writing; they must be experienced if they are to become part of the experimenter's intuitive sense of what makes for good experimental design. All of us have experienced the difference between listening to an instructor telling us how to run a regression analysis in principle and running a regression analysis in practice. The difference between running an experiment in principle and in practice is even greater because an experiment has more moving parts, more uncontrollable elements: human beings as opposed to data that just sits quietly on a shelf. Running experiments requires practice practice practice. In my opinion, it takes several thousand dollars worth (more if anything) of failed experiments for an experimenter to acquire the intuition and practice that will allow him or her to gather decent data.

I continued to run my experiments by hand even after I moved up on the learning curve for reasons that are idiosyncratic to my theory. The agents in my model are "collectively connected" in two ways. First, each agent has partial information that is relevant to other agents' voting decisions. Second, one agent cannot unilaterally overturn the status quo; she needs other agents to go along. The informational externality along with the collective nature of the voting decision raises the possibility that my results are sensitive to whether the experimental design is common knowledge – does each participant understand the game she is playing, and does each participant believe that all other participants understand the game they are playing, and does each participant believe that all other participants believe that all participants understand the game they are playing, . . . ?

Because I was concerned about establishing common knowledge, I conducted my experiments in a very transparent and public way. I designed the urns and chips so that all participants could observe each other getting their private information, with the information itself remaining private. All aspects of the formal incentive structure of the game were publicly observable or publicly verified. Whether I truly established common knowledge is not something that can be known for sure. But my design compares favorably, I think, with a computer design where a computer screen informs the participants about their private information and about the distribution from which other participants' private information is drawn – but the participants cannot observe or verify whether the computer is telling the truth. Granted, in practice most participants will believe the computer, but in

the case of my theory, the issue is not only whether the participants believe the computer but also whether they believe that everybody else believes that everybody believes . . . . the computer. In a game-theoretic model, the agents' beliefs are an integral part of the model, and depending on the model it may be more or less important to control the participants' beliefs as best as possible and establish common knowledge to the extent possible.

The fourth step in running experiments is to write instructions. Here I cannot emphasize enough how important it is to write and rewrite the instructions and rewrite them yet again. From a rational choice perspective, that is, in a world with no cognitive limitations, the clarity of the description of a game is unimportant. In practice, clear instructions tend to reduce noise in the results. Just like newly minted PhDs tend to submit unfinished drafts to refereed journals, blissfully unaware that they are about fifty rewrites away from anything remotely resembling a publishable manuscript, a first-time experimenter has a tendency to think his or her instructions are clear when they are not.

Another judgment call is whether to write the instructions in abstract form ("the GGW urn pays one dollar, the WWG urn zero dollars") or with a substantive application in mind ("if the incumbent government is competent, you get one dollar, if it is incompetent, zero dollars"). Experimental economists tend to prefer abstraction, psychologists concrete applications. Once again, each approach has its upsides and downsides. Abstract formulations are less likely to trigger stray thoughts and emotions that might influence people's behavior; on the other hand, participants usually find it easier to develop an intuitive grasp of the game they are playing when it is presented in the context of a concrete application.

Experimental economists tend to implement an abstract formulation with the help of urns, chips, bingo cages, colored balls, coin tosses – the works, and they tend to go out of their way to make the incentive structure credible. In fact, these paraphernalia, as well as the attempts to establish credibility, may well trigger associations with the way magicians operate and thus make participants suspicious about deception – participants who otherwise go through life mostly taking things at their face value and not expecting to be deceived.

The punchline for an experimenter is: abstract or concrete, you can't win either way. More generally, there are no perfect solutions in experimental design, and the goal has to be to find a good trade-off between imperfect solutions.

Another important question is how to design the monetary stakes. Higher monetary stakes have the potential to translate into less noisy data (because competing non-monetary motivations play a relatively smaller role) and improve the performance of game theory (whether this is true is, of course, an empirical question, though economists often behave as if it must be true a priori). With a limited budget, the stakes should be designed so that at the margin the payoff between making good and bad decisions (as measured by the theory to be tested) makes a big difference in monetary payoffs. This guiding principle is generally useful but it is not very helpful for experiments involving collective action. Here, the participants' monetary payoffs depend on each others' actions in possibly very complicated and subtle ways, and the probability that one action will make a difference for the outcome may be very small. Indeed, I would say that the point of testing theories of collective action in the laboratory is to find out how people actually behave in situations with very low-powered incentives.

One important decision an experimenter faces is which of the comparative statics of her theory to test. A good theory generates many different comparative statics. Gathering experimental data is a very costly enterprise. For a given budget there is always a tradeoff between varying parameters across experiments so as to test as many comparative statics as possible and repeating experiments so as to improve the levels of significance.

As an unexperienced experimenter, I ended up changing my experimental design over time, not to test my theoretical comparative statics but to get a sense of how different experimental designs affect the results. First, I was initially very concerned that my monetary stakes were too small. So I increased the stakes in two experiments (2/24/97 and 3/3/97). I ended up spending a lot of money with no obvious difference in the results. Second, I initially attempted to establish credibility and promote common knowledge by randomly drawing a monitor from the group of participants and having him or her verify that the incentive structure as it was laid out in the instructions was in fact the true incentive structure (2/3/97 and 2/24/97). I got the impression that having a monitor didn't do much of anything – the participants were just as willing (or unwilling) to believe me. I also experimented with computer-generated random numbers instead of coin tosses (2/24/97). I subsequently moved to a simpler experimental design that allowed the participants to believe their own eyes. Third, one experiment that lasted twenty rounds (2/24/97) taught me how important it is to keep things constant over the course of the experiment – participants are overwhelmed if too much changes and they need to continuously adjust their decision rules. For the first ten rounds, I kept reshuffling the participants into new groups so as to dampen repeated game effects; for the second ten rounds, I allowed the participants to remain in the same groups so as to encourage learning effects. I found out that allowing the participants' decision rules to equilibrate,

or to best respond to each other and eventually settle down, was more important than avoiding repeated game effects (which don't have much of a chance if the decision rules move around too much).

One of the biggest problems I faced was how to prevent informational leakage. I might think that the GGW and WWG urns are outwardly indistinguishable, but creases in the plastic and dirty smudges allow an attentive participant to take a perfect informational cue from the urn itself than an imperfect cue from the color of the chips she privately observes. In one experiment, one row of participants faced outward into a window. Cardboard boards separated them sideways, but I had not thought of putting boards between the participants and the window, nor had I thought of drawing the curtains. When it got dark outside in the course of the experiment, the participants' mirror images showed up in the window, and they ended up signaling their private information to each other. The fact that they found it worthwhile to do so suggests that they did understand the point of the experiment (information pooling), but it also meant that they were playing a different game than the one I was testing.

Again and again, one or another participants would figure out how to circumvent the "rules of the game," the information structure being particularly vulnerable; I would then work hard to redesign the experiment; in the next experiment, some other participant would find another loophole in the design. After many experiments, I finally succeeded in running experiments with no informational leakage (as far as I know), but to this day I approach each experiment with the expectation that it is vulnerable to "spoilers." On a more positive note, I have come to appreciate the capacity of human beings to break out of, or find loopholes in, formal incentive structures. What motivates participants to undermine the rules of the game is not only the prospect of making more money; there is the boredom of sitting in an experiment, there is the glee at outfoxing a professor, . . . How difficult it is to control participants in an experiment tells us something, I think, about the vulnerability of real-world formal incentive structures to the creative machinations of human beings.

With more experience, I now have less waste due to bad (in hindsight) experimental design. There is no substitute for experience. Rational choice theories do not nail down human behavior with sufficient accuracy. There are many tradeoffs and judgment calls in experimental practice, and there is a limit as to how much one can theorize in advance how a specific experimental design will work out.

Experimental Results

Taken literally, my model of dynamic informational cascades was "rejected" in the laboratory. I use quotation marks because I have not yet resolved the question of what is the appropriate null hypothesis. If the null is "everybody abstains all the time," then the null is rejected in favor of my model. To complicate matters, everybody abstaining all the time is in fact one – admittedly trivial – equilibrium of my model.

Qualitatively, the model performs quite well in the laboratory. Most participants most of the time appear to use simple decision rules that "make sense" in the context of the model; that is, their individual-specific stakes, their private information, and the past history of the cascade appear to influence their decisions to cast costly votes or abstain in a way that is more or less consistent with the model. The model also predicts whether a cascade is prone to a status quo bias or an anti-status quo bias and whether errors of type I or II are likely to occur (such errors occur when the status quo is overturned even though a fully informed majority would have preferred to see it maintained, and vice versa). These predictions are also fulfilled, admittedly in a rather messy way (so much for my hope that I would be able to run away from the messiness of the real world by going to the laboratory).

A full description of the experimental results exceeds the purpose and scope of this article. Let me mention one surprising result. Overall, the participants in the three-person groups tended to cast fewer costly votes than did the participants in the fifteen-person groups even though the free-rider problem is obviously more severe in the latter. The three-person groups tended to end up in one of two situations. Some of them succeeded in coordinating on something resembling the full-information equilibrium. Others miscoordinated, or they suffered under the abstentions of a "spoiler" (now defined as a participant who abstains throughout an experiment independent of her individual-specific stakes and private information); either way, they tended to get dragged down to the zero-information equilibrium where everybody abstains all the time. The fifteen-person groups, on the other hand, rarely experienced a round with total abstentions. There was always the odd one or two participants who would cast a costly vote, for no obvious reason or against apparently great odds that their votes would do much of anything; sometimes other participants would go along, sometimes they would not, in which case one or the other active participant would give up – but then someone else would jump into the fray.

I do not yet fully understand why the behavioral dynamics of the small N and large N groups are so different. My hunch is that in the small N case people respond to each other as individuals, using the aggregate number of votes to figure out each others' individual decision rules, whereas in the large N case people take a "statistical mechanics"

(Durlauf, 1986) approach, interpreting the aggregate number of votes as a noisy signal of the underlying distribution of private information. This insight does not come out of the theory – on the contrary, the theory suggests that the free rider problem, which tends to dampen participation in costly collective action, is more severe in large groups than in small groups.

Rational Choice and Collective Action

We need to ask ourselves, I think, whether we should expect the rational choice paradigm to work well for collective action given the low-powered incentives involved. The rational choice paradigm is usually justified with reference to learning and selection effects. But it is typical of collective action that people do not get detailed information about how everyone else behaved as a function of their stakes or private information; people typically only get very simple aggregate statistics (how many people took part in a demonstration, how many people voted for one candidate or another). For large N, people typically get very weak feedback telling them that the decision rules mapping their preferences, information, and past history into a prescription to behave one way or the other are counterproductive. Similarly, when the probability is very small that one individual's behavior makes a difference for the collective outcome, the selection effects rewarding productive strategies and punishing counterproductive strategies have little bite.

Laboratory experiments (along with casual introspection and real-world data) tell us that rational choice analysis with its narrow focus on the instrumental motivations underlying political participation may be blind to important empirical features of collective action: people's participation decisions are shaped by ethical considerations, emotional factors, and cognitive limitations. The rational choice approach can be usefully extended to allow for ethical and expressive motivations identified by economists and political scientists (Rabin, 1993; Schuessler, 1996) and cognitive limitations identified by evolutionary psychologists (Cosmides and Tooby, 1994).

There is, of course, nothing new in "explaining" political participation with reference to altruism, social norms, social embeddedness, selective incentives, political leadership, cognitive biases, psychological benefits of participation or costs of nonparticipation, and the like. The ad hocism inherent in such approaches is obviously unattractive. Instead, evolutionary selection effects may serve to place restrictions on the scholarly practice of invoking ad hoc assumptions about variables in the utility function or cognitive biases.

In my current work (Lohmann, 1988), I examine whether and how the way people reason about political participation is hardwired by evolutionary selection effects. In my model, agents are heterogeneous "types": their utility functions differ with respect to some parameter, or they differ with regard to some cognitive bias. Because of their type, some agents contribute more real resources to the public good than do others. They get higher utility payoffs from doing so, but lower payoffs as measured in real resources. Whether a specific type "survives" depends on her real resources. Moreover, because the distribution of types affects the provision of the public good, it influences the rate of turnover ("deaths") in the society as a whole. This model serves as a starting point to examine how selection effects shape the distribution of agent types over time and to relate the patterns generated by evolutionary simulations to the patterns observed in empirical data on collective action.

More generally, laboratory experiments allow us to identify the conditions under which people deviate from the normative prescriptions of game theory and to figure out how people deal with situations where game theory makes no prediction. Computational models have the potential to "rationalize" such deviations in the context of a larger rationality – evolution. The work of Van Huyck, Battalio, and Beil (1990) and Arifovic (1996) is an excellent example of how laboratory experiments and computational models can complement each other in spelling out how and why people cope with the strategic uncertainty posed by multiple equilibria. Theirs is an example which I hope others will follow.

References

Jasmina Arifovic. 1996. "Strategic Uncertainty and the Genetic Algorithm Adaptation." Mimeograph. Simon Fraser University.

Cosmides, Leda, and John Tooby. 1994. "Better Than Rational: Evolutionary Psychology and the Invisible Hand." *American Economic Review Papers and Proceedings* 84: 327-332.

Durlauf, Steven N. 1996. "Statistical Mechanics Approaches to Socioeconomic Behavior." Mimeograph. University of Wisconsin at Madison and Santa Fe Institute.

Latour, Bruno. 1987. *Science in Action.* Cambridge: Harvard University Press.

Lohmann, Susanne. 1994. "Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989-1991." World Politics 47: 42-101.

Lohmann, Susanne. 1997a. "Dynamic Informational Cascades." Mimeograph. University of California, Los Angeles.

Lohmann, Susanne. 1997b. "Stand Up and Be Counted: An Informational Rationale for the Power in Numbers." Mimeograph. University of California, Los Angeles.

Lohmann, Susanne. 1998. "Do People Have a Taste For Helping Others or Do They Have a Taste for Punishing Others for Not Helping Others?" Mimeograph. University of California, Los Angeles.

Lohmann, Susanne. Forthcoming. "I Know You Know She Knows We Know You Know They Know: Common Knowledge and the Unpredictability of Informational Cascades." In *Political Complexity: Nonlinear Models of Politics*, ed. Diana Richards. Ann Arbor: University of Michigan Press.

Lohmann, Susanne. Work in Progress. "Mass Action and the East German Revolution." Manuscript. University of California, Los Angeles.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83: 1281-1302.

Schuessler, Alexander. 1996. "Expressive Motivation and Electoral Mass Participation." Mimeograph. New York University.

Van Huyck, John B., Raymond C. Battalio, Richard O. Beil. 1990. "Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure." *American Economic Review* 80: 234-249.

Useful Background Readings

Camerer, Colin F. 1997. "Progress in Behavioral Game Theory." *Journal of Economic Perspectives* 11: 167-179.

Conlisk, John. 1996. "Why Bounded Rationality." *Journal of Economic Literature* 34: 669-700.

Cosmides, Leda, and John Tooby. 1992. "Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions From the Literature on Judgment Under Uncertainty." Mimeograph. University of California, Santa Barbara.

Douglas D., and Charles A. Holt. 1993. *Experimental Economics*. Princeton: Princeton University Press.

Friedman, Daniel, and Shyam Sunder. 1994. *Experimental Methods*. Cambridge: Cambridge University Press.

Kagel, John H., and Alvin E. Roth. 1995. *The Handbook of Experimental Economics*. Princeton: Princeton University Press.

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

Selten, Reinhard. 1989. "Evolution, Learning, and Economic Behavior." Nancy L. Schwartz Memorial Lecture. Northwestern University.

Posterior Distributions from Model Averaging: A Clarification[1]

Larry M. Bartels
Princeton University
lbartels@wws.princeton.edu

In a recent *AJPS* article on "Specification Uncertainty and Model Averaging" (Bartels 1997), I described a procedure for combining statistical results derived from alternative model specifications. The aim of "model averaging" is to allow simultaneously for statistical uncertainty about parameter values within any given model and for specification uncertainty arising from the fact that different more or less plausible models may produce different estimates of the same underlying parameters.

In the course of addressing a critique of my article presented by Robert Erikson, Gerald Wright, and John McIver at the 1997 Political Methodology Conference, I realized that my original exposition provided a misleading characterization of the posterior distributions resulting from model averaging. My aim in the present note is to clarify the shape of these posterior distributions. The resulting revision has no significant impact on the thrust or conclusions of my original article, but provides a better intuitive understanding of how model averaging works, as well as a more precise algorithm for hypothesis testing using the model-averaging procedure.

In the Bayesian framework from which model averaging is derived, the main product of any statistical analysis is a posterior distribution reflecting uncertain beliefs about a parameter (or parameters) after reconciling prior beliefs with observable data. In general, this posterior distribution may have a quite complicated form. However, in a simple multiple regression framework with diffuse prior beliefs represented by normal distributions and normally distributed data, the Bayesian posterior distribution is simply a normal distribution with mean and standard deviation equal to the parameter estimate and standard error from a classical regression analysis (Leamer 1978, 77-79). In this sense, the classical parameter estimate and its standard error provide a good representation of the posterior beliefs of a Bayesian

[1]**I am grateful to Robert Erikson, Gerald Wright, and John McIver for stimulating the clarification reported here, and to Renée Smith for a helpful reading of an earlier draft.**

with very weak prior beliefs about the value of the parameter in question.

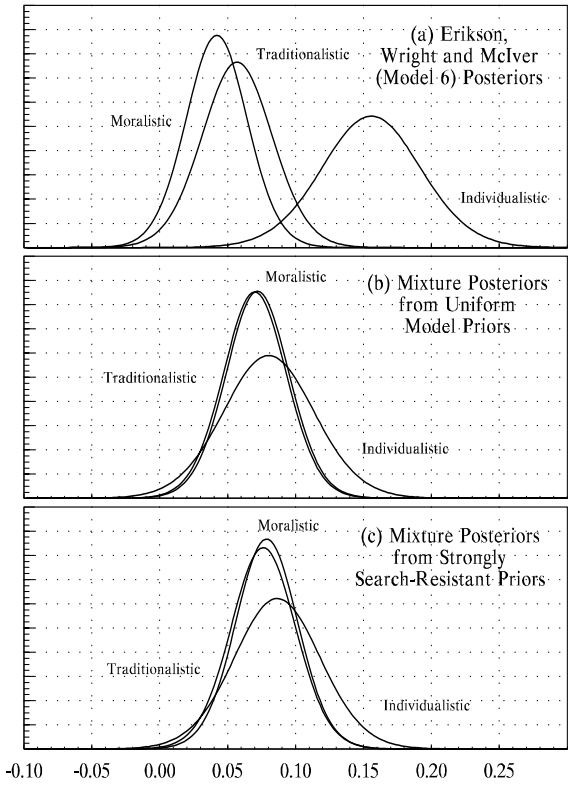
In the model-averaging procedure, the posterior distribution of each parameter is a mixture distribution reflecting uncertainty both within and across models. If the conditional posterior distributions representing our uncertain beliefs about a parameter under each alternative model are normal distributions – as they will be under the assumptions set out in my article – then the unconditional posterior distribution resulting from model averaging will be a *mixture* of normal distributions, but not itself a normal distribution.

My original presentation obscured this fact in two ways: by using the means and standard deviations of posterior distributions from model averaging to compute "t- ratios," and by picturing the mixture posteriors in my figures as normal distributions. I explained in a footnote (Bartels 1997, 654) that I reported t-ratios for the model mixture coefficients "for descriptive purposes only," and that "[t]he t-distribution will not be a good approximation for the actual posterior distribution of any coefficient set to zero with certainty in models which get significant posterior weight, since the posterior will be a mixture of a t-distribution and a spike at zero" reflecting the posterior weight of models in which the corresponding parameter is omitted.

In fact, the posterior distribution of each parameter under the assumptions in my article is a mixture of normal distributions (the Bayesian analogs to the t-distributions arising in classical regression theory), and this mixture of normal distributions will not, in general, be a normal distribution even when there are no spikes produced by models in which the corresponding parameter is omitted. Whether a normal distribution with the same mean and standard deviation will be a "good approximation" to this mixture of normal distributions depends upon the distinctiveness of the posterior distributions arising from the various separate models under consideration – and upon the use to which the approximation will be put.

The distinction between a normal distribution and a mixture of normal distributions may be clarified by comparing my original Figure 3, which presented alternative posterior distributions for the effects of state opinion on policy outcomes in Erikson, Wright, and McIver's (1993) Individualistic, Moralistic, and Traditionalistic states, with the revised Figure 3 presented here. In each figure, panel (a) shows the posterior distributions of the relevant effects implied by Erikson, Wright, and McIver's (1993) preferred regression model, panel (b) shows the posterior distributions implied by model averaging with uniform priors attached to six distinct regression models, and panel c shows the posterior distributions implied by model averaging with strongly "search-resistant" priors attached to the six regression models. The means and standard deviations of the distributions are identical in every case, but panels (b) and (c) of the revised figure show the correct mixture distributions rather than the normal approximations to those mixture distributions presented in the original figure.

Original Figure 3: Alternative Posteriors for Culture-Specific Effects of State Opinion

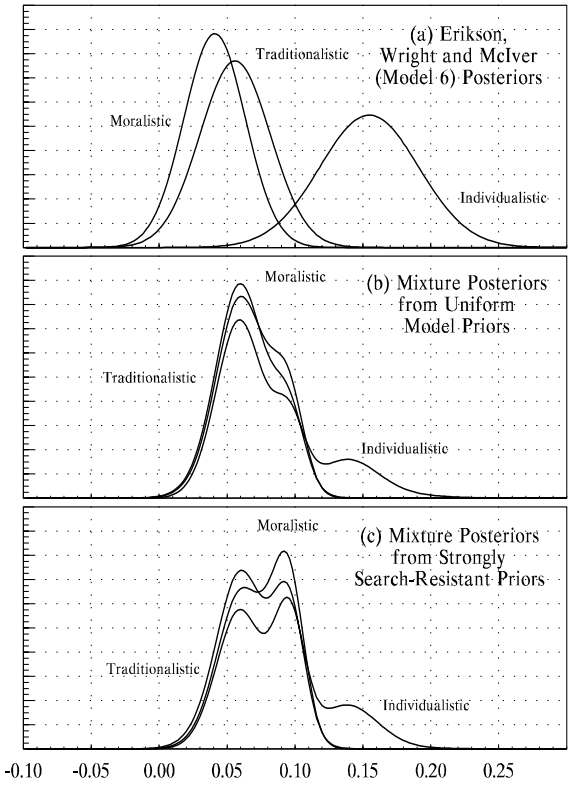


The mixture distributions in panels (b) and (c) of the revised Figure 3 display distinct modes reflecting the different parameter estimates from the alternative regression models reported by Erikson, Wright, and McIver (1993). In the posterior distributions for the Individualistic states, one of these modes implies a noticeably stronger effect of public opinion on state policy than in Moralistic or Traditionalistic states. This mode represents the posterior weight attached to the two models (including Erikson, Wright, and McIver's Model 6 from panel (a)) in which policy outcomes appear to be especially responsive to public opinion in Individualistic states. Together, these two models receive about 14 percent of the posterior weight in both mixture posteriors.

The multimodal posterior distributions illustrated in the revised Figure 3 provide a somewhat clearer sense of how the model-averaging procedure reconciles the results from alternative model specifications. In essence, the posterior distribution produced by model averaging is simply a

combination of the posterior distributions for the same parameter from the various separate models, each weighted by the posterior probability associated with that model. If only one model receives any weight (as in panel (a) of Figure 3), the resulting posterior essentially replicates the classical result. However, when more than one model receives appreciable posterior weight – and when the parameter estimates implied by those models are significantly different – the resulting posterior may not be well approximated by a normal distribution with the same mean and standard deviation.

Revised Figure 3: Alternative Posteriors for Culture-Specific Effects of State Opinion



In this particular example, the difference is of little substantive import: both the original and revised posterior distributions portray "a virtual overlap of the three distinct distributions" of estimated effects for Individualistic, Moralistic, and Traditionalistic states (Bartels 1997, 665) once model uncertainty is taken into account, rather than the sharp distinction suggested by Erikson, Wright, and McIver's most preferred model (represented in panel (a)).

It is worth noting, however, that, even in this case, specific inferences may be sensitive to the precise shape of the posterior distribution. For example, Erikson, Wright, and McIver (1997) have noted that classical tests of the null hypothesis that a given parameter is equal to zero may be misleading if based simply on the "t-statistic" computed by dividing the posterior mixture coefficient by its standard deviation in the usual way, since the actual posterior distribution is not a t-distribution. More generally, formal hypothesis tests should be based upon mixture posteriors like the one in the revised Figure 3 presented here, rather than upon normal approximations like the one in my original Figure 3. The correct calculation of a tail probability for the mixture posterior requires computing the corresponding tail probabilities in each of the separate distributions for the various alternative models and then adding up these separate tail probabilities, each weighted by the posterior probability associated with the corresponding model.

Of course, for many purposes reporting the mean and standard deviation of a mixture posterior will be sufficient to convey the implications of a model-averaging analysis. Nevertheless, analysts and readers alike should be aware that the actual posterior distribution that these statistics summarize has a somewhat different form than in corresponding analyses based upon a single statistical model.

R eferences

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41:2, 641-674.

Erikson, Robert S., Gerald C. Wright, and John P. McIver. 1993. *Statehouse Democracy: Public Opinion and Policy in the American States*. Cambridge: Cambridge University Press.

Erikson, Robert S., Gerald C. Wright, and John P. McIver. 1997. "Too Many Variables? A Comment on Bartels' Model-Averaging Proposal." Presented at the 1997 Political Methodology Conference, June 1997, Columbus, Ohio.

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

Program for the 1998 Summer Methodology Meeting - UCSD

Thursday Morning

Larry M. Bartels, Princeton "Panel Attrition and Panel Conditioning in American National Election Studies," **Discussant:** Charles Franklin, Wisconsin.

Gary King, James Honaker, Anne Joseph, and Kenneth Scheve, Harvard "Listwise Deletion is Evil: What to do about Missing Data in Political Science," **Discussant:** Chris Achen, Michigan.

Thursday Afternoon

Michael C. Herron, Northwestern; "Two-Candidate Elections, Ecological Inference, and Implied Spatial Voting," **Discussant:** John Londregan, UCLA.

Invited Talk

Halbert White, Department of Economics, UCSD, "A Reality Check for Data Snooping".

Thursday evening

Graduate Student Poster Sessions

Friday Morning

Panel A

Gregory J. Wawro, Columbia "A Dynamic Panel Analysis of Campaign Contributions," **Discussant:** Michael Alvarez, Caltech.

Nolan McCarty, Columbia, and Lawrence Rothenberg, Rochester, "The Time to Give: PAC Motivations and Electoral Timing," **Discussant:** Michael Bailey, Georgetown.

Panel B

John Brehm, Duke, "Comparison of Methods of Analysis of Time Allocation," **Discussant:** Douglas Rivers, Stanford.

Philip Paolino, Texas, "Two Approaches to Maximum Likelihood Estimation of the Beta Distribution," **Discussant:** Jonathan Nagler, UC Riverside.

Friday Afternoon

Panel A

John Freeman, Minnesota "Democracy and Exchange Rates: An Experimental Study," **Discussant:** Suzanna De Boef, Penn State.

Walter Mebane, Cornell "Rational Expectations Coordinating Voting in American Presidential and House Elections," **Discussant:** Curtis S. Signorino, Rochester.

Panel B

Dan Wood, Texas A&M, "Estimating Time Varying Parameter Models with Flexible Least Squares," **Discussant:** Nathaniel Beck, UC San Diego.

Langche Zeng, George Washington, "Data Analysis with Neutral Network Models," **Discussant:** Philip A. Schrodt Kansas.

Saturday Morning

Brad Jones, Micheal E. Sobel, Arizona "Modeling Ideological Trends using Adjacent Category Logit Models for Ordinal Scales with Midpoints," **Discussant:** Henry E. Brady, UC Berkeley.

Burt L. Monroe, Indiana, "Bias, Responsiveness, Swing, Majoritarianism and Disproportionality: Sense and Nonsense in the Analysis of Seats-Votes and Other Representation Relationships," **Discussant:** Mohan Penubarti, UCLA.

Saturday Afternoon

Patrick Bryant and John Williams, Indiana "A State-Space Approach to Event Counts: Structural Models and Monte Carlo Evidence," **Discussant:** James Stimson, North Carolina.

Simon Jackman, Stanford, "Bridging Statistical 'No Man's Land': Time Series Models for Qualitative Data," **Discussant:** Renee M. Smith Rochester.

1998 APSA Short Courses

At the 1997 APSA meeting the Political Methodology Society sponsored two short courses: one on discrete choice models, and one on ecological inference. These courses were very successful, and three courses are being sponsored at the 1998 APSA meeting in Boston. The courses will take place on Wednesday, September 2 - the day before the meeting begins. The course on Bayesian Analysis takes place in the afternoon, from 1pm - 4pm. The Event History Course takes place in the morning from 9:30 - 12:30; and the Time-Series-Cross-Section Data course takes place in the afternoon from 1pm - 4pm. Thus it is possible to combine both of the latter two courses. Registration for each course is $10. The 1997 courses were attended by people ranging from graduate students to full professors. If you have questions about what the courses will cover, feel free to contact the instructors - each of who's email address is given below.

Course 1: Bayesian Analysis

– Simon Jackman

Over the last ten years, the statistics literature has been abuzz with the development of Bayesian simulation methods for estimation and inference. These methods — Markov chain Monte Carlo (MCMC) methods — hold great potential for social scientists, which is only now starting to be tapped. The short course will provide an introduction to MCMC methods, starting with the EM algorithm, first developed for imputing value on missing data. Gibbs sampling

— the dominant MCMC technique — is based on a generalization of the EM algorithm. As the title of the short course suggests, these methods have a Bayesian heritage, though can be usefully applied by researchers without a strong commitment to Bayesian inference.

Participants will be shown how MCMC methods advance our understanding of a wide range of empirical issues in American politics, comparative politics and international relations. These include: the determinants of presidential approval and macropartisanship, models from comparative political economy (e.g., welfare state expenditures, union density), the analysis of aggregate data and ecological inference, heterogeneity and ambivilence underlying American public opinion, the determinants of international conflict, models of vote choice (multinomial probit). Course participants will be introduced to the **BUGS** program (Bayesian inference Using Gibbs Sampling), and supplied with data and code for the examples listed above.

Contact: Simon Jackman, *jackman@stanford.edu.*

Course 2: Event History Models in American Politics, Comparative Analysis, and International Relations

– Janet Box-Steffensmeier
– Brad Jones

Events-oriented data are prevalant in political analysis. Whether one is examining the onset of military conflict, the termination of a political coalition, or the ending of a legislative career, the issue of timing of an event–that is, when some event occurs–is implicitly important. And while events-oriented data are common in American politics, comparative analysis, and international relations, methods for analyzing events history are less well understood. In this short course, we will provide an intermediate introduction to event history methods. Topics considered in this short course will include the following:

- Why standard regression models are inadequate in the face of event history data.
- Nonparametric and parametric estimation of event history models (including a look at the Cox proportional hazard model and the Weibull distribution).
- Inclusion of time-varying covariates (TVCs) and interpretation of TVCs in the context of event history data.
- Special problems that arise with event history data (including a consideration of duration dependency, simultaneity of TVCs, and heterogeneity).

For this short course, we assume no prior knowledge of event history methods; however, we will assume a basic understanding of the class linear regression model as well as a "conceptual" understanding of maximum likelihood estimation.

Contact: Brad Jones, *bsjones@u.arizona.edu.*

Course 3: Taking Time and Space Seriously (Particularly in Comparative Politics and International Relations): A Short Course on Time-Series–Cross-Section Data

– Neal Beck

An introduction to time-series–cross-sectional (tscs) data with particular stress on applications to data in comparative politics and international relations. Topics to be covered include:

- The advantages of tscs data
- The setup of tscs data
- Estimating tscs data (OLS, GLS, PCSE's)
- The dynamics of tscs data
- TSCS data with a binary dependent variable

The course is accesible to anyone who has had a course in regression or is familiar with OLS.

This course may be taken in conjunction with "Event History Models in American Politics, Comparative Analysis, and International Relations."

Contact: Neal Beck, *nbeck@weber.ucsd.edu.*

The Political Methodologist
Department of Political Science – 073
University of California
Riverside, CA 92521-0422

The Political Methodologist is the newsletter of the Political Methodology Section of the American Political Science Association. We gratefully acknowledge the support of the College of Humanities, Arts, and Social Sciences of the University of California, Riverside 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 (212 483-2512) to join the section. Dues are $8.00 per year.

Submissions to *TPM* are welcome. Articles should be sent to the editor by e-mail (mohan@ucla.edu) if possible. Alternatively, submissions can be made on diskette as plain ascii files sent to Mohan Penubarti, Department of Political Science, Bunche Hall, UCLA, Los Angeles CA 90095-1472. LaTeX format files are especially encouraged. See the *TPM* web-site [http://wizard.ucr.edu/polmeth/tpm/] for the latest information.

TPM was produced using LaTeX on a PC running Linux and a Hewlett-Packard workstation running HP-UX.