Public Administration Research and Practice: A Methodological Manifesto[1]

Jeff Gill
California Polytechnic State University

Kenneth J. Meier
Texas A&M University

[1] Paper prepared for presentation at the Fifth National Public Management Research Conference, George Bush School of Government and Public Service, Texas A&M University, December 3-4, 1999.

1 Introduction

Public Administration, in Dwight Waldo's terms, has become a profession. With that achievement, examining the methodological infrastructure of the profession is merited. Every profession rests on an infrastructure of research and research methods on which the profession's practitioners base their day-to-day activities. This essay will argue that public administration has ignored its technical side and that given the types of problems dealt with by both academics and practitioners, a serious upgrading of methodological skills is needed. We hope to provide a road map, useful to both methodologists and non-methodologists, for developing those skills.

To date, public administration has relied heavily on related disciplines for its methodological tools. The utility of such a strategy is open to question on several grounds. First, the disciplines we borrow from are interested in different types of questions, and the methods they adopt are driven by those specific questions. Political science with a few exceptions has moved away from the consideration of questions of organization and management. A recent survey by Wise (1998) found that of the top 25 ranked Ph.D. programs in political science, only four permitted a doctoral student to offer a major field in public administration. Political scientists have focused their methods on questions of voting behavior and have developed their skills accordingly. While such methods have some utility to scholars of public administration, especially those who do survey research, they do not grapple with important questions of measurement and estimation unique to public administration. Economics offers even fewer opportunities for the transfer of methods. The discipline has moved strongly away from empirical analysis to a focus on theory. While applied econometricians abound and find favor except in their own discipline, the econometric approach is driven almost completely by economic problems. As a result, it relies heavily on strong distributional assumptions and generally downplays issues of measurement, robustness, and visual presentation.

Second, public administration lacks the financial resources to follow the practice of medicine and co-opt other disciplines into developing its methods. The large medical research infrastructure provided by the National Institutes of Health provides ample incentives for scholars in numerous disciplines to apply their interests to questions useful to the medical profession. Public administration, in contrast, is a stepchild. Study after study has found that public administration research attracts little funding from major foundations or

agencies (Perry and Kramer 1986). A recent assessment by Brintnall (1998) confirmed this in regard to the National Science Foundation. Even by Brintnall's liberal definition of "public administration," the allocation of funding to the area was minuscule.

Third, the practical side of public administration adds additional demands on methods that make borrowing from other disciplines less promising. If one is to use quantitative methods to prescribe policy change, then one clearly needs better methodological skills that those used by most other social sciences. If a political scientist makes a major error in his or her study of the 1992 election, it matters little. Clinton still wins. If a public administration scholar commits a major error in analyzing a education program, it can have major implications simply because it could influence public policy. Much of welfare policy seems to be driven by the idea of welfare migration (the movement of individuals to collect higher benefits); an idea resulting from a single flawed study (Peterson and Rom 1989, Berry, Fording, Hanson 1999).

Fourth, if the practice side of public administration cautions against the adoption of the best practices of political science methods, it totally rejects the approach of economics. Economic models work well by assuming away the real world and constructing artificial markets to correct all policy ills. The practitioner in contrast would rather have a model that failed dismally in theory but proved relatively accurate in the real world.

The solution, we feel, is for public administration to invest heavily in developing its own methods. Even if it relies heavily on other fields to develop the specific quantitative techniques used, it needs to select those techniques based on key public administration research questions rather than those advocated by the discipline in question. There are six key methodological developments that we feel are essential to public administration: independent data sources, avoiding the null hypothesis significance test, increased use of time series analysis, the adoption of Bayesian methods, the transition from estimation techniques to optimization techniques (SWAT), and the use of the general linear model (GLM). We conclude by discussing the incentives needed to further these developments and note other potential areas where breakthroughs are possible.

2 Independent Data Archives

Public administration differs from other fields and professions in that it lacks a set of core data sets that comprise the basic data infrastructure for the field. Political scientists have the National Election Studies, the Correlates of War, easily obtained congressional and state legislative data sets; economists have the Citicorp data set with its national time series, data from the Bureau of Labor Statistics; demographers and sociologists have the Panel Study on Income Dynamics, the General Social Survey, and the High School and Beyond data sets among others.

The absence of a collection of core data sets has both disadvantages and advantages. First, teaching research in public administration is more difficult simply because students do not have ready access to data sets concerned with management and administrative questions. In the social sciences, first semester graduate students can conduct actual research projects by just using data that can easily be accessed through the archives at the University of Michigan or similar places.

Second, students of public administration lack a common exposure to sets of data that would provide a notion of what core research questions might be. The result is less interaction among researchers and, thus, greater difficulty in generating a cumulative body of research.

While the disadvantages of a lack of core data sets are limiting, if one gets beyond the problem, it becomes an opportunity. Unlike other social scientists, public administration scholars are not trapped by data sets and funneled into studying only certain questions. How the National Election Studies have structured what is considered political science is a classic case of path dependence in this regard. The absence of core data sets means that public administration scholars expect to gather their own data and, thus, are more sensitive to issues of measurement reliability and validity. The absence has also contributed to more eclectic use of data sets including government documents, elite interviews, archival research, and the merger of multiple data sources in addition to survey research.

Given our long experience in working with nonstandard data sets, creating an infrastructure of core public administration data sets should not adversely affect the current advantages. At the same time, an archive of core data sets would have substantial payoffs in graduate education. Class time could be

conserved for actual analysis; instructors would be familiar with the data sets the students were using; texts could tie problem sets directly to these data sets; replication exercises could be used.

The idea behind public administration core data sets and an accessible archive is to supplement currently available data not to replace them. In particular, many government data sets need supplemented because the data were gathered with specific (perhaps non-research) purposes in mind. In some cases, data might actually be biased or collected only when it serves the interests of the agencies involved. Two examples serve as illustration. In 1982 the Reagan administration eliminated the tracking system for family planning funding. Since that time no official data are available on number of individuals served, the types of services offered, the number of service locations and other crucial policy information. In 1994 after the first four years of the Milwaukee School Choice Experiment, Governor Tommy Thompson eliminated all requirements to collect data on students involved in the experiment, thus ending any useful policy information that could be gained from the program.

In other cases private data sources illustrate problems with public databases. Individuals who work in the area of abortion policy have long recognized that abortion incidence data published by the Centers for Disease Control and Prevention are woefully incomplete. Data collected by the Alan Guttmacher Institute, in contrast, are far more complete and generally of more use from a policy perspective. Unfortunately, private organizations rarely have the resources to collect data on government programs over an extended period of time.

The best way to create a core set of public administration data sets is with the establishment of a data archive where current data sets could be stored. Data sets that are not archived are frequently lost; a prominent example is the American Federal Executive study of 10,000 high level civil servants, 10,000 military officers, and 10,000 civilian elites conducted by W. Lloyd Warner and colleagues (1959). Only recently was the lost 1955 Thorndike-Hagen study of 5,085 Air Force veterans found again and converted from virtually unusable 9-track tape. Much data gathered only 10 to 15 years ago is rapidly becoming unreadable as new data retrieval technologies are developed.

Public administration organizations and journals need to encourage all individuals to archive their data, preferably at the same archive. Even if the data are not made available immediately, they could well produce some additional knowledge at some point in the future when they are available. In addition, having a group of public administration scholars define what might be a set of core of variables to be gathered, either cross-sectionally or longitudinal would be useful. As an illustration, a time series of federal government employment, expenditure and structural data could be used by a variety of different scholars to address key questions of public administration.

Another pervasive problem relates to generalizability and data aggregation in public administration data sets. All too often state and local government data are combined (Lewis and Nice 1994), despite important and substantively relevant differences in the two levels of government. Conversely, generalizing from a single state to national effects is also a common pathology in the public administration literature. Given the heterogeneity of the U.S. states (ignoring international public administration issues for the moment), this seems particularly unwise. In addition, it is not uncommon to see state level inferences made from dangerously small subgroups. For example, Newman (1994) looks at the effects of occupational segregation by sex using a sample of Florida public administrators, but the sample contains only 29 women. The sex ratio itself is certainly a statistically valid measure, but to infer policy output characteristics from a sample this small is risky.

Another data problem plagues public administration. It is what Tufte (1977) calls: "Data Dumping, or Putting Together of Statistical Compilations with a Shovel." The problem lies in the tendency for data collections to amass simply because the data are available. Availability and importance are easily confused in the large quantity of data currently produced from government sources. In addition the material called "data" by many furnishing agencies really is not. There is an appalling tendency for analysis in the form of simple tables and histograms to be labeled as data when they are really a product of such data.

A final data set problem bears mentioning. There are a number of studies that use sensitive, often government archived, data to study public administration questions. Most often these data sets are from personnel files where protection of anonymity is legally and ethically warranted. Rather than sanitize the file and place it in the public domain for all interested scholars; however, authors

will frequently embargo the data. So readers of the resulting publication must take the authors' claims on faith. (c.f. Miller, Kerr, and Reid 1999). This practice clearly retards the growth of empirical public administration scholarship because it not only weakens any inferential claim made by authors, but it also precludes any replication study.

While we have portrayed a fairly grim picture of the current state of data archiving in public administration, we would be remiss not to point out some successes. The Center for Urban Policy Research at Rutgers University has archived and made freely available The State of the Nation's Cities: A Comprehensive Database on American Cities and Suburbs at http://www.policy.rutgers.edu/cupr/sonc.htm. The Center for Presidential Studies at the Bush School of Government and Public Service at Texas A&M University has developed an ongoing project to archive presidential data for general access at **http://www-bushschool.tamu.edu/CPS/archive/index.html**. A similar archive is being created for public management at http://www-bushschool.tamu.edu/pubman/. The University of California, San Diego provides a very useful index of links to social science data sets which contains a surprisingly high proportion of links to data sets of interest to public administration researchers: **http://odwin.ucsd.edu/cgi-bin/easysearch2?search=getdata&file=/data/data.html&print=notitle&header=/header/data.header.** In addition, federal government provision of data has improved dramatically over the last few years (despite our previous comments). Useful federally funded access points include the Government Information Sharing Project (**http://govinfo.kerr.orst.edu/index.html**), the Federal Inter-agency Council on Statistical Policy Fedstats site (**http://www.fedstats.gov/**), the Government Information Exchange which includes the Federal Yellow Pages (**http://www.info.gov/**), the Government Information Locator Service (GILS) (**http://www.access.gpo.gov/sudocs/gils/gils.html**), and the very useful starting point at Fedworld (**http://www.fedworld.gov/locator.htm**).

3 The Flawed Practice of Null Hypothesis Significance Testing (or: Stars Are Stupid)

The currently employed method of hypothesis testing employed in public administration, and every other social science field, is logically and interpretively flawed in the deepest possible sense (Gill 1999). Even though this approach provides misleading conclusions from statistical results, null hypothesis significance testing (NHST) has dominated the reporting of empirical results in the social sciences for over fifty years. Because public

administration methodology is a young subfield, the investment in this bankrupt enterprise is relatively mild compared to other disciplines. We, therefore, hope that it is abandoned before taking hold the way it has in related areas.

3.1 How the NHST Works

The ubiquitous NHST is a synthesis of the Fisher test of significance and the Neyman-Pearson hypothesis test. In this procedure, two hypotheses are posited: a null or restricted hypothesis (H_0) competing with an alternative or research hypothesis (H_1) each describing complementary notions about some social or administrative phenomenon. The research hypothesis is the model describing the researcher's assertion about some underlying aspect of the data, and operationalizes this assertion through a statement about some parameter, β. In the most basic case, described in every introductory text, null hypothesis asserts that $\beta = 0$ and a complementary research hypothesis asserts that $\beta \neq 0$.

A test statistic (T), some function of β and the data, is calculated and compared with its known distribution under the assumption that H_0 is true. The test procedure assigns one of two decisions, D_0 or D_1, to all possible values in the sample space of T, which correspond to supporting either H_0 or H_1 respectively. The p-value is equal to the area in the tail (or tails) of the assumed distribution under H_0 which start at the point designated by the placement of T and continuing away from the expected value to infinity.

If a predetermined α level has been specified, then H_0 is rejected for p-values less than α, otherwise the p-value itself is reported as evidence for H_1. Thus decision D_1 is made if the test statistic is sufficiently atypical given the distribution under the assumption that H_0 is true.

This NHST process is a synthesis of two highly influential but incompatible schools of thought in modern statistics. Fisher's (1925) procedure produces significance levels from the data whereas Neyman and Pearson (1933a, 1933b, 1936) posit a test-oriented decision process which confirms or rejects hypotheses at a priori specified levels. The NHST test attempts to blend these two approaches. In Fisher hypothesis testing, no explicit complementary hypothesis to H_0 is identified. The p-value that results from the model and the data is evaluated as the strength of the evidence for the research hypothesis. There is no notion of the power of the test, the probability of correctly

rejecting H_0. Nor is there an overt decision in favor of H_1. Conversely, Neyman-Pearson tests identify complementary hypotheses: Θ_A and Θ_B in which rejection of one implies acceptance of the other, and this rejection is based on a predetermined α level. Thus there is an overt decision in this process.

Neyman and Pearson's hypothesis test defines the significance level (α = 0.10, 0.05, 0.01,...) a priori as a function of the *test* (i.e. before even looking at the data), whereas Fisher's test of significance provides the significance level afterwards as a function of the *data*. The current paradigm in the social sciences straddles these two approaches by pretending to select α a priori, but actually using p-values or worse yet, asterisks next to test statistics indicating ranges of p-values ("stars"), to evaluate the strength of the evidence. This allows inclusion of the alternate hypothesis but avoids the often difficult search for more powerful tests.

The NHST is also an attempt to reconcile the two differing perspectives on how the hypotheses are defined. It adopts the Neyman-Pearson convention of two explicitly stated rival hypotheses, but one is always labeled as the null hypothesis as in the Fisher test. In some introductory texts the null hypothesis is presented only as a null relationship: $\beta = 0$ (i.e. no effect), whereas Fisher really intended the null hypothesis simply as something to be nullified. The synthesized test partially uses the Neyman-Pearson decision process except that failing to reject the null hypothesis is incorrectly treated as a quasi-decision: modest support for the null hypothesis assertion. There is also confusion in the NHST about p-values and long-run probabilities. Since the p-value, or range of p-values indicated by stars, is not set a priori, it is not the long-run probability of making a Type I error but is typically treated as such.

3.2 The Inverse Probability Problem

The most common interpretive problem with the NHST is a misunderstanding of the order of the conditional probability. Many public administration academics and practitioners incorrectly believe that the smaller the p-value is, the greater the probability that the null hypothesis is false: that the NHST produces $P(H_0|D)$, the probability of H_0 being true given the observed data D. In fact, the NHST first posits H_0 as true and then asks, what is the probability of observing these or more extreme data? This is unambiguously $P(D|H_0)$. A more

desirable test would be one that produces $P(H_0|D)$ because this would facilitate a search for the hypothesis with the greatest probability of being true given the observed data. Bayes law shows the difference between these two unequal probabilities:

$$P(H_0|D) = \frac{P(H_0)}{P(D)} P(D|H_0). \qquad (1)$$

The two quantities, $P(D|H_0)$ and $P(H_0|D)$, are equal only if $P(H_0) = P(D)$, and the probability of this equality is exactly zero.

Jeffreys (1961) observed that using p-values as decision criteria in this way is backward in its reasoning: "a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred." Others have noted "a p-value of 0.05 essentially does not provide any evidence against the null hypothesis" (Berger, Boukai, and Wang 1997).

3.3 The Decision Problem With the NHST

Unlike most social scientists, public administrators are experts in decision making. Many have had formal classes in decision making so they should understand the implications of the NHST in the decision-making process. As described above, we use the test statistic to make one of two decisions: D_0 or D_1. In fact, only these two decisions, or actions, are allowable. So it is technically incorrect to make statements from the NHST such as "provides modest evidence." Despite this, many authors confuse the decision process with the strength of evidence. That is, the NHST interpretation of hypothesis testing confuses inference and decision making since it "does not allow for the costs of possible wrong actions to be taken into account in any precise way" (Barnett 1973).

Worse yet, the researcher reaches one of these decisions in the NHST where the cost of being wrong is completely exogenous to the decision process. Thus the cost of being wrong is completely abstract at the time of the decision. A more reasonable approach is to assign a loss function to each of the two decisions. This is a real-valued function that explicitly provides a loss for decision *i* given β^* is the true parameter value: $L(\beta^*, H_i)$. So from this we can build a decision rule that codifies some criteria that the researcher might have: minimize the maximum loss, minimize squared errors, and many others. The quality

of the decision rule is generally judged by a corresponding risk function. This is simply the average loss across decisions by using this decision rule for a given value of β^*. So good decision rules have lower risk functions than known alternatives.

It should be clear from this brief discussion that a major component of hypothesis testing and decision theory is omitted from the NHST. One need not add all of the formality implied by the last paragraph, but it is important to realize that the NHST decision does not include the subsequent consequences.

3.4 The Model Selection Problem

Typically when a public administration model result with a NHST is reported, it is presented as if only two models were ever considered: the null hypothesis and the provided research hypothesis. The quality of a research finding is then solely judged by the ability to reject the single complementary null hypothesis with a sufficiently low p-value. However, during the development of the reported model many differing alternate mixes of independent variables are often tested. This is called the "illusion of theory confirmation" (Greenwald 1975, Lindsay 1995) because the NHST is presented as evidence of the exclusivity of explanation of this single research hypothesis. Statistics are reported in published form from the final model specification as if many other model specifications tried in the development never existed and that this last model is produced from a fully controlled experiment (Leamer 1978, p.4). The NHST test thus provides an infinitely strong bias in favor of a single research hypothesis against an infinite number of other hypotheses (Rozeboom 1960, Lehmann 1986, p.68, Popper 1968, p.113).

Two completely plausible and statistically significant models can lead to entirely different conclusions about the substantive question of interest using the exact same data (Raftery 1995). In many cases the decision criteria that led to the final model are based at least in part on intermediate significance levels and that the significance levels reported in the final published work have very different interpretations than the significance levels in intermediate models (Leamer 1978, Miller 1990, Raftery 1995). The worst example of this is the use of stepwise regression (sometimes called "unwise regression", see King 1986), which replaces theory with a mechanical process driven exclusively by ordered mixing of covariates according to residual sum of square minimization. The problem with stepwise regression, besides of course the odious way in which

it dispenses with actual theories about the phenomenon of interest, is that it does not even necessarily reach the best conclusion as specified by its own flawed criteria.

The process of mining through a data set trying to find statistically significant results without a theoretical foundation is always a poor research strategy. If there are twenty potential relationships all of which are actually not present in a given population, then we are still likely to find, by sampling induced chance, at least one to be statistically significant at the $\alpha = 0.05$ level. As evidence to support this claim, use any statistical software package to create a dataset with 20 completely independent random variables in columns of length approximately 100. Now look at the correlation matrix between these 20 variables. If this process is repeated only a few times you will find that about 5% of the time a correlation coefficient tested with $\alpha = 0.05$ is determined to be significant. So it is almost guaranteed that data-mining leads to statistically significant results which make some claim about a public administration question subject to a Type I error with probability one. This fallacy is like the "lottery paradox" which refers to situation in which nearly every large lottery winner has some unusual attribute which appears to make their prior probability of winning very low. However, this search for some unusual attribute takes place after the person has won the lottery. So no matter how unusual or unlikely the attribute is in the population, the probability that this person would win is 1 since they have already won (journalists are inevitably deceived here).

Data mining is related to the classic "file drawer problem." In a hypothetical world suppose that there are no relationships to be found, but there are dedicated scientists working away and publishing their "significant" findings. In this world the journals would be filled with studies that made Type I errors and the scientists' file drawers would be filled with unpublished studies that were correct. Fortunately we live in a world where there are non-zero effects, and there is evidence that the file drawer problem is not pervasive (Rosenthal 1979, Rosenthal and Rubin 1988).

Rosenthal and Rubin (1978) look at all 345 annual publications in psychology and notice that the mean z-value from empirical studies was 1.22. Working backwards they calculate that there would have to be 65,123 studies in the file drawers for that year with a zero mean z-value in order to conclude that the published

work was entirely due to sampling effects. Their finding is promising (provided that psychology is representative of the social sciences in general and we really do not have massive file cabinets) because it implies that a reasonable number of correct decisions are made with regard to hypothesis testing and publication.

3.5 The Probabilistic Modus Tollens Problem

The basis of the NHST rests on the logical argument of modus tollens (denying the consequent). The basic strategy is to make an assumption, observe some real-world event, and then check the consistency of the assumption given this observation. The modus tollens syllogism works like this:

- If H_0 is true then the data will follow an expected pattern,
- the data do not follow the expected pattern,
- therefore H_0 is false.

The problem with the application of this logic to hypothesis testing is that the certainty statements above are replaced with probabilistic statements, causing the logic of modus tollens to fail. To see this, reword the logic above in the following way:

- If H_0 is true then the data are highly likely to follow an expected pattern,
- the data do not follow the expected pattern,
- therefore H_0 is highly unlikely.

This logic seems plausible. However, it is a fallacy to assert that obtaining data that is atypical under a given assumption implies that the assumption is likely false: almost a contradiction of the null hypothesis does not imply that the null hypothesis almost false (Falk and Greenbaum 1995). For example:

- If an agency has a "glass ceiling" then it is highly unlikely to see women as senior managers
- the agency has a female senior manager
- Therefore it is highly unlikely that the agency has a "glass ceiling".

From this simple example and the resulting absurdity it is easy to see that if the P(Female Senior Manager|Glass Ceiling) is low (the p-value), it does not imply that P(Glass Ceiling|Female Senior Manager) is also low. In other words, the NHST does not protect the user from logical inconsistencies that arise from

ill-defined, non-mutually exclusive competing sets (Cohen 1994, Pollard and Richardson 1987).

3.6 The Significance Through Sample Size Problem

There are two important misinterpretations about the impact of sample size in the NHST. First is the common belief that statistical significance in a large sample study implies real world importance. A NHST based on a large dataset almost always results in statistical significance in the form of low p-values (Leamer 1978, Macdonald 1997, Oakes 1986, Raftery 1995). This is a concern in public administration research since it is incorrect to infer that some subfields have greater legitimacy just because the corresponding data sets tend to produce smaller p-values: "a prejudice against the null" (Greenwald 1975, p.1). For instance, contrast the expected results from a study of federal employees in which data on tens of thousands of individuals is analyzed versus a study of the budget of national security agencies. Clearly in the latter world, sample sizes are vastly smaller.

The correct interpretation is that as the sample size increases we are able to progressively distinguish smaller population *effect sizes*. This is the extent to which some measured phenomenon exists in the population. Finding population effect sizes is actually the central purpose of social science research since any hypothesized difference can be found to be statistically significant given a sufficiently large sample. Public administration researchers are really more interested in the relative magnitude of effects (program success, budget changes, legislative support, levels of representativeness, etc.), and making merely binary decisions about the existence of an effect is not particularly informative. However, the NHST deflects us into an obsession with the strength of this binary decision measured through p-values.

The second misinterpretation is that for a given, observed p-value in a study which rejects the null hypothesis, larger sample sizes imply more reliable results. This is false because once the p-value is observed, the sample size is already taken into account (conditioned on). Two studies that reject the null with the same p-value are equally likely to make a Type I error even if they have dramatically different sample sizes. This mistake actually results from a poor understanding of Type II errors. Two studies which *fail* to reject the null hypothesis and are identical in every way except sample size are different

qualitatively: the test with the larger sample size is less likely to make a Type II error.

To explore this misconception Rosenthal and Gaito (1963) asked 19 university researchers to express their "degree of belief in research findings" for two studies: one with a sample size of 10 and another with a sample size of 100, but having the same p-value. Most of the respondents had more confidence in the results of the test with the sample size of 100 despite the identical p-values. Rosenthal and Gaito inferred that researchers include the sample size effect on the probability of a Type II error as a qualitative measure of null hypothesis significance testing. This is a dangerous qualitative measure since if the null hypothesis is rejected, the probability of making a Type II error is zero.

In a more recent study, Wilkerson and Olson (1997) asked 52 psychology graduate students to evaluate two tests which report p-values of 0.05 and are identical in every way except that one has a sample size of 25 and the other has a sample size of 250. The graduate students were asked which test had the greatest probability of making a Type I error. Only 6 out of the 52 correctly observed that the two tests have an identical probability of falsely rejecting the null hypothesis.

3.7 The Arbitrariness of Alpha Problem

Fisher constructed the first significance level tables and, therefore, personally established the conventional rejection level thresholds. The majority of Fisher’s work was applied to agricultural experiments and biometrics. Familiar contributions included the analysis of variance of various treatments on plant growth: sunlight, fertilizer, and soil conditions. While there may be some amount of fertilizer in empirical public administration research, it is clearly not of the kind envisioned by Fisher.

On what basis do we decide that p = 0.051 is inadequate evidence for rejection but p = 0.049 is perfectly adequate to reject? Distinctions at this level rely upon the assumption that there is virtually no measurement error, an assumption that no informed social scientist would ever be willing to defend. No published work in any social science field provides a theoretical basis for these thresholds. While it is convenient to say "one time out of twenty" or "one time out of a hundred," but it is no less convenient to say "one time out of fifty" or "one time out of 25" (i.e. p = 0.02 and p = 0.04). Fisher’s justification

rests on no established scientific principle. Instead he believed that these levels represented some standard convention in human thought. So we make fundamental substantive conclusions in public administration research based on an early twentieth century bio-statistician's intuition about how people think about probabilistic evidence?

3.8 The Accepting the Null Hypothesis Problem

When we fail to reject the null hypothesis, we cannot conclude that the null is true because doing so does not rule out an infinite number of other competing research hypotheses. The NHST is asymmetric: if the test statistic is sufficiently atypical assuming the null hypothesis then the null hypothesis is rejected, but if the test statistic is insufficiently atypical assuming the null hypothesis then the null hypothesis is not accepted. This has been called a double standard: H_1 is held innocent until proven guilty, and H_0 is held guilty until proven innocent (Rozeboom 1960).

There are two problems that develop as a result of asymmetry. The first is a misinterpretation of the asymmetry to assert that finding a non-statistically significant difference or effect is evidence that it is equal to zero or is nearly zero. Regarding the impact of this acceptance error Schmidt (1996, p.126) asserts that this: "belief held by many researchers is the most devastating of all to the research enterprise." This acceptance of the null hypothesis is damaging because it inhibits the exploration of competing research hypotheses. The second problem pertains to the correct interpretation of failing to reject the null hypotheses. Failing to reject the null hypothesis essentially provides almost no information about the state of the world. It simply means that given the evidence at hand one cannot make an assertion about some relationship: all you can conclude is that you cannot conclude that the null was false (Cohen 1962).

3.9 An Easy Solution for Public Administration Research: Confidence Intervals

Confidence intervals are an alternative to the NHST that provide the same information and more. In addition, confidence intervals do not require a contrived decision. In the simplest case a confidence interval is constructed by taking a point estimate of some underlying population parameter, $\hat{\beta}$, and enveloping it with a probability structure that extends a number of standard errors of $\hat{\beta}$ (σ_β) in both directions: $[\hat{\beta} - \sigma_\beta k_\alpha : \hat{\beta} + \sigma_\beta k_\alpha]$. The familiar

textbook example is produced when we believe that the sampling distribution of $\hat{\beta}$ is gaussian-normal and therefore apply $k_{\alpha=0.95,2\text{-tail}} = 1.96$.

Confidence sets, the more general case of confidence intervals where the region is not required to be contiguous, are estimates of some parameter β in which the uncertainty is expressed as a range of alternate values and a probability of coverage. Credible sets, Bayesian set estimates described below, measure the probability that the parameter is in the interval rather than the probability that the interval covers the true parameter. With confidence sets the set itself is the random quantity and the unknown parameter is fixed. So we cannot state that with any produced confidence set there is a known probability that the unknown parameter is contained, we have to say that we are $(1 - \alpha)$% confident that this set covers the true parameter value.

Confidence intervals and the NHST present the same information: a linear regression coefficient with a $1 - \alpha$ confidence interval bounded away from zero is functionally identical to a NHST rejecting at $p \leq \alpha$ the hypothesis that the coefficient equals zero. However, confidence intervals have a superior feature: as the sample size increases the size of the interval decreases, correctly expressing our increased certainty about the parameter of interest. This is analogous to the correct interpretation of increasing statistical power in a NHST as sample size increases. Most misunderstandings about sample size as a quality measure in the NHST stem from a poor understanding of Type II errors. Since there is no Type II error in confidence intervals, there is less potential for such confusion.

Unfortunately the confidence level of the interval is subject to the same arbitrary interpretations as α levels. Therefore confidence intervals require the same cautions with regard to sample size interpretations and unsupported conventions about α levels.

An interesting and unusual example and one of the earliest explicit works in social science methodology is Sir Isaac Newton’s last manuscript, *The Chronology of Ancient Kingdom’s Amended* (1728). Newton estimates the mean length of the reign ancient kings from biblical times to his present era in order to refute a current claim that the average interval was between 35 and 40 years. In this analysis he determines that the "medium" range is "about eighteen or twenty

years a-piece." What makes this result interesting is Newton's use of an interval estimate rather than a point estimate for the length of reigns. Stigler (1977) shows that while Newton did not explicitly use the sampling distribution of the mean or a maximum likelihood estimate, both of these techniques produce intervals almost identical to Newton's, demonstrating that he distinguished between the distribution of the mean versus the distribution of the data. Newton, therefore, presented his results as an interval around the mean reign in order to convey both the measure of centrality and some uncertainty that he felt about this measure.

Effective alternatives to the NHST exist that require only modest changes in empirical methodology: confidence intervals, Bayesian estimation, and meta-analysis. Confidence intervals are readily supplied by even the simplest of statistical computing packages, and require little effort to interpret. Bayesian estimation eliminates many of the pathologies described, albeit with a greater setup cost (see below). Meta-analysis, looking across multiple independent studies, offers the potential benefit of integrating and analyzing a wider scope of work on some administrative or policy question (Gill 1999).

4 Time Series Analysis

Time series analysis is well suited to before-after program evaluation designs, and thus has received a modest amount of use in public administration. The adoption of new programs can be viewed as natural experiments when program outcomes are measured for a period of time before the new program is adopted. Lewis-Beck and Alford (1980) used interrupted time series analysis to determine the impact of various coal mine safety laws; the results of the analysis were then used to generate some theoretical propositions about when laws have their intended impact. Morgan and Pelissero (1982) addressed whether or not reformed city government structures resulted in changes in public policy outputs and concluded that they did not. Durant, Legge, and Moussios (1998) assessed the changes in British Telcom's behavior after privatization using an ARIMA model to determine if it engaged in the predicted competitive behavior. Other examples of time series analysis include Wood and Waterman (1994), Wood and Peake (1998), Meier (1980) and countless more.

Despite the relatively frequent use of time series, we feel that the potential for use in public administration has barely scratched the surface. Several interesting theoretical questions can be addressed through the use of more

advanced time series techniques including the responsiveness of inertial systems, the "causality" among variables, the use of pooling to overcome data problems, and the incorporation of long run equilibrium techniques.

4.1 Inertial Systems

Both organizations and policy can be characterized as inertial systems-present outputs are a function of past results. Organizations quite clearly are regularized patterns of interaction that attempt to impose stability on a policy process. Stability also appears to be valued in policy because it allows one to avoid reinventing the wheel each year and attempt incremental changes in program rules.

Theoretically, an inertial system is best represented by the following time series:

$$\mathbf{Y}_t = \alpha + \beta_1 \mathbf{Y}_{t-1} + \beta_2 \mathbf{X} + \varepsilon \qquad (2)$$

where **Y** is an output vector of interest and **X** is either a policy intervention vector or a matrix of relevant environmental variables.

This autoregression equation is superior to the common use of a trend line or counter variable (see Lewis-Beck and Alford 1980) because it does not impose a strict linear pattern on the system and it generates results that are more likely to be consistent with organizations and policy programs.[2] The impact of any change in the system such as a new program (**X**) affects the system's outputs over a period of time rather than all at once. The coefficient β_2 is the impact of a one unit change in **X** on **Y** for the first year of the change. Because **Y** at time *t* then also affects **Y** at time *t + 1* and in future years, the initial change in **X** continues to affect **Y** through the lagged dependent variable. In the second year of the program, this impact is $\beta_2 x \beta_1$. For subsequent years, the impact continues but at a declining rate, forming what is called a distributive lag model (Pyndick and Rubinfeld 1991).

The distributive lag formulation, we feel, is consistent with our theories of organization. Small initial changes can over time result in fairly large long run impacts. The Social Security reform debates are contemporary illustrations

of how very small initial savings can generate substantial long run results for a policy system. The distributive lag also implies that exogenous impacts (the **X** variables) must compete with standard operating procedures and bureaucratic routines to influence the organization's policy outputs.

The autoregressive model creates some addition methodological problems that need to be addressed. With a lagged dependent variable both the traditional Durbin-Watson test and the Box-Ljung Test for serial correlation are biased (Maddala 1992). The appropriate test is the LaGrange Multiplier test based on the Gauss-Newton regression (Davidson and McKinnon 1993). One regresses the residuals from the equation on all independent variables plus multiple lags of the residuals (substituting zeros in for missing values created by the lagging procedure). The R-square of this equation is multiplied by the number of cases to obtain the test statistic which is chi-square distributed with the degrees of freedom equal to the number of lagged residuals. The test is a joint test that is sensitive to both serial correlation and moving average problems. The LaGrange Multiplier test is appropriate whether or not the dependent variable is lagged so it should be the general test for serial correlation in any time series.

4.2 "Causality"

Many time series questions in public administration are questions of causality-do representative bureaucracies generate different policies, or do organizations with different policies also decide to become representative bureaucracies? Does school system bureaucracy cause poor performance among students, or does poor performance among students cause a school system to add programs and thus bureaucracy (Smith and Meier 1995)? Cause as used in public administration and social science has a precise meaning that reflects the famed "chicken and egg problem" of which came first (Thurman and Fisher 1988). By causality we generally mean that when **X** changes **Y** will subsequently change but when **Y** changes, **X** will not change in any predictable way. Causality is thus, a temporal ordering that can be assessed statistically.

The simplest way to determine a "causal" linkage is with several cases that are measured at two different times. Using the above illustration, let us assume a group of 100 school districts with measures of student performance and bureaucracy for two different years. The logic of the test is that if

[2] For an example of this approach using the management of organizations and networks, see O'Toole and Meier (1999). For a policy related approach using the area of agricultural credit, see Meier, Polinard, and Wrinkle (1999).

bureaucracy causes student performance to drop that bureaucracy at time 1 will be negatively correlated with student performance at time 2 (as performance comes into congruence with bureaucracy) when one controls for the level of student performance at time 1. At the same time, bureaucracy because it is not caused by student performance will change randomly from time 1 to time 2 with respect to student performance. Student performance at time 1, therefore, will be uncorrelated with bureaucracy at time 2 when controlling for bureaucracy at time 1. The following equations set up this panel test of causality:

$$\mathbf{Y}_t = \alpha + \beta_1\mathbf{X}_{t-1} + \beta_2\mathbf{Y}_{t-1} + \varepsilon \qquad (3)$$

$$\mathbf{X}_t = \alpha + \beta_3\mathbf{X}_{t-1} + \beta_4\mathbf{Y}_{t-1} + \varepsilon \qquad (4)$$

A hypothesis of **Y** causes **X** is tested by a significant coefficient for **Y** in equation (4). A hypothesis that **X** causes **Y** is tested by a significant coefficient for **X** in equation (3). Of course, variables can be reciprocally related, and both hypotheses can be supported (or both rejected). For illustrations of the technique see Meier and Smith (1994) relating representative bureaucracies to political representation or Smith and Meier (1995) relating public school performance to private school enrollments.

Panel analysis assumes that the causal period, that is the time that it takes **X** to affect **Y**, is equal to the time from *t - 1* to *t*. The technique also assumes that one lag is sufficient to incorporate all the pass history of a variable. The former assumption can be assessed by extending the time lag (if the question is whether or not the lag is long enough); the latter assumption can be directly dealt with by using either Granger (1969) causality tests or vector autoregression.

Granger causality tests involve two time series of data with sufficient points to allow analysis (N > 30). Rather than assuming a single lag of a variable is significant, Granger causality incorporates multiple lags. The logic behind Granger causality is that if **X** causes **Y**, then one can predict **Y** with multiple lags of **X** even when multiple lags of **Y** are included in the equation. Statistically one estimates equation (5), the restricted equation, and equation (6), the unrestricted equation, as follows:

$$\mathbf{Y}_t = \alpha + \sum\beta_j\mathbf{X}_j + \varepsilon \qquad (5)$$

$$\mathbf{Y}_t = \alpha + \sum\beta_j\mathbf{X}_j + \sum\beta_i\mathbf{X}_i + \varepsilon \qquad (6)$$

The appropriate test is a joint f-test to determine if all the coefficients for the **X** variables are equal to zero (and thus add nothing to the level of prediction). The test for whether or not **Y** causes **X** is done analogously by setting up a similar set of equations. A full discussion of Granger Causality can be found in Freeman (1983); for an application in public administration, Wood (1992) uses Granger causality tests to sort out the various possible relationship in environmental regulatory enforcement.

If Granger causality can be thought of as a bivariate test of causality, that is a test between two time series, then vector autoregression (VAR) is merely the multivariate analogue to Granger causality. Granger causality can indeed produced spurious results if key variables are omitted from the test (see Meier and Smith 1994). Vector autoregression allows for the additional variables to be included as blocks of variables in the unrestricted equation. Each block is assessed as a block to determine if it should be included in the system of equations. A good illustration of the use of VAR is Krause (1996) who examined securities regulation by assessing the linkages between Congressional preferences, Presidential preferences, and agency behavior. Krause (1996) finds that agency behavior exerts a great deal of influence on both Congress and the Presidency, a finding that challenges much of the literature spawned by the empirical principal-agent approach.

4.3 Pooled Analysis

Pooled time series is often an option when the analyst has some data over time but not necessarily enough data to run a full time series. It is also used for relatively long time series when one thinks that the processes are similar in different environments (e.g., the impact of welfare policies in various Western democracies). Pooled data sets allow the to analyst assess more total cases and thus circumvent problems such as collinearity and too few data points relative to the parameters that need to be estimated.

Because pooled time series analysis requires attention to serial correlation, heteroscedasticity and cross-correlations (Stimson 1985, Hsaio 1986), it is relatively difficult to do correctly. Data must be organized correctly and this organization will vary depending on the software being used. Without correct organization, all the diagnostics used to assess data problems are at best

misleading and most likely wrong. Preprogrammed software is available. STATA and SAS both have relatively adequate pooled packages for panels that have more cross sections than time points (what are affectionately know as wide but shallow pools). For pools that have more time points than cross sections (narrow but deep pools), Eviews has an excellent package; SAS also has a program but it is unsupported and may generate misleading results (Beck and Katz 1996).

Several examples of pooled time series exist in the relevant literature. Keiser and Meier (1996) used a pooled model to examine the impact on several federal laws on state-level child support collection policies. Keiser and Soss (1998) used a similar technique to probe the determinants of discretionary welfare decisions. Hedge, Menzel and Williams (1988) have examined surface mining regulation at the state level using data for several years for the approximately 20 states that operate their own regulatory programs (as opposed to those that permit the federal government to regulate surface mining).

4.4 Stationarity and Cointegration: Much Ado About Nothing?

A relatively common problem in time series analysis is that the data series is not stationary. A stationary series is one that has a constant mean and variance over time. Series that are trending or contain dramatic breaks are generally nonstationary series. The basic problem caused by stationarity is that two variables that are uncorrelated but both nonstationary (integrated is another term) are susceptible to spurious correlation (Granger and Newbold 1976). De Boef and Granato (1999) provide simulation results to show that the common 0.05 level of significance is rejected as much as 50 percent of the time with two stationary but unrelated series.

The solution to nonstationary data is to difference the data (subtract last year's value from this year's) until the data are stationary.[3] Differencing data creates its own problems. Over-differencing (differencing when it is not needed) can induce a moving average problem in the data. Differencing also limits the analyst to examining short term impacts, that is, change in **X** this year and its impact on the change in **Y** this year. Many programs and policy, as noted above, have substantial long run dynamics; such processes can be modeled with

[3] There are numerous tests for stationarity. Unfortunately, all of the tests have weaknesses so that some judgment is required to conclude that a series is stationary.

differenced data but only introducing relatively complex terms (e.g., polynomial distributed lags).

Cointegration is a method that examines both long run and short run processes for any data that are integrated at the same level (that is, need to be differenced the same number of times). The methodology of cointegration is not complex (see Clarke and Stewart 1994 for an example), but work in political science has consistently misapplied the technique in an atheoretical manner. In economic theory, cointegration exists only when two series are in long run equilibrium (e.g., supply and demand must be in long run equilibrium). Political science lacks theories that predict long run equilibrium for most of the cases where cointegration has been applied (see Ostrom and Smith 1992, Beck 1992, Williams 1992, Durr 1993). Public administration, however, could well have processes that are theoretically cointegrated. In states with balanced budget amendments, expenditures and revenues must be in long run equilibrium. Within a given policy area with authority shared between the federal and state governments, a long run equilibrium is also possible. If theory development progresses to where it demands the estimating of relationships that are in long run equilibrium, then the techniques of cointegration are available. Until theory demands the technique, however, cointegration might well simply be much ado about nothing (see Maddala 1992, Li and Maddala 1996).

5 Likelihood and Bayesian Methods

The practice of developing empirical models in public administration research is often very straightforward. Frequently this involves no more than specifying a linear model or developing some crosstabs. Sometimes, however, a more complex structure is needed to model some social or administrative phenomenon. Therefore we often develop parameterized non-linear models in which we want an estimate of unknown quantities and a measure of reliability of that estimate: in short the values that have maximum probability of being true given the observed data. Unfortunately the NHST and its inverse probability problem get in the way here. This problem was discussed in Section 3.2, but no solution was given. In this section we provide two ways to make inferences that avoid the inverse probability problem.

Suppose we consider the collected data as fixed to us. Suppose further that the parametric form of the proposed model is a component of the hypothesis. Neither of these statements are in the least bit controversial because at some point in

the analysis we must decide to condition on the existing data, and except for exploratory studies, we have to make *some* statement about the probability structure that produced the data (even in non-parametric approaches there are such assumptions made). Likelihood and Bayesian methods are similar in that they start with these two suppositions and develop estimates of the unknown parameters in the parametric model.

5.1 Maximum Likelihood

Maximum likelihood estimation finesses the inverse probability problem discussed in Section 3.2 by substituting the unbounded notion of likelihood for the bounded definition of probability. This is done by starting with Bayes Law:

$$P(\beta|D,H_0) = \frac{P(\beta,H_0)}{P(D,H_0)} P(D|\beta,H_0) \qquad (7)$$

where β is the unknown parameter of interest, D is the collected data, and H_0 is the null hypothesis including the assumed specification of the parametric form of the data. The key is to treat $P(\beta,H_0)/P(D,H_0)$ as an unknown function of the data independent of $P(D|\beta,H_0)$. This allows us to use: $L(\beta|D,H_0) \propto P(D|\beta,H_0)$. Since the data are now fixed and the hypotheses stated, we get different values of the likelihood function only by inserting different values of the unknown parameter, β.

The likelihood function, $L(\beta|D,H_0)$, is similar to the desired but unavailable inverse probability, $P(D|\beta,H_0)$, in that it facilitates testing alternate values of β, to find a most probable value: $\hat{\beta}$. Because the likelihood function is no longer bounded by zero and one, it is now important only relative to other likelihood functions based on differing values of β.

Interest is generally in obtaining the "maximum likelihood" estimate of β. This is the value of the unconstrained (here) and unknown parameter, β, which provides the maximum value of the likelihood function, $L(\beta|D,H_0)$. This value of β, denoted $\hat{\beta}$, is the most likely to have generated the data given H_0 expressed through a specific parametric form relative to other possible values in the sample space of β.

Maximum likelihood estimation was introduced to modern statistics by Fisher (1925), but its origins are generally credited to Gauss (see Stigler (1986, p.141) or Brenner-Golomb (1993, p.299) for interesting discussions). It can safely be said that maximum likelihood estimate is the workhorse of twentieth century statistics. Excellent mathematical statistics discussions can be found in Casella and Berger (1990, Chapter 7), Rohatgi (1976, Chapter 8), Hogg and Craig (1978, Chapter 6), and Stuart and Ord (1994, Chapter 8). Basic econometric texts generally cover maximum likelihood estimation in some detail: Greene (1999, Chapter 4), Gujarati (1995, Chapter 4), Maddala (1992, Chapter 3), Judge et al. (1982, Chapter 6). Advanced econometric texts with excellent technical discussions include Amemiya (1985, Chapter 4), and Schmidt (1976, Chapter 3).

5.2 Bayes

The Bayesian approach addresses the NHST inverse probability problem discussed in Section 3.2 by making distributional assumptions about the unconditional distribution of the parameter, β, prior to observing the data, $P(\beta|H_0)$. This sometimes called a "subjective probability" because it comes from the researcher's prior knowledge or estimate before looking at the data. Often there are strong theoretical justifications for this prior probability such as information from previous studies, suppositions in the relevant literature, and the researchers own expertise. Although public administration scholars rarely use Bayesian methods, the ability to incorporate expertise through the priors makes it a method that should be used frequent in the area.

Assigning a prior probability on the unknown parameters is very useful because it provides a means of integrating out β to solve for the previously unknown $P(D|H_0)$ in (7):

$$P(D,H_0) = \int_{\beta\in\beta} P(D|\beta,H_0)P(\beta,H_0)d\beta. \qquad (8)$$

This allows us to avoid the inverse probability problem because we now have a value for the denominator in (7). With this construct, the conditional (posterior) distribution of β is proportional to the likelihood times the prior:

$$P(\beta|D,H_0) \propto P(D|\beta,H_0)P(\beta|H_0) \qquad (9)$$

since the data are assumed fixed and therefore have no relevant distribution: $P(D,H_0)$. The Bayesian data analysis approach is focused around getting this estimate of the distribution of the unknown parameter value "post" to the data.

Bayesians are not concerned with getting a specific point estimate of β because it is assumed to have a distribution (the posterior calculated above) rather than being fixed but unknown. The Bayesian focus is on describing the shape and characteristics of this posterior distribution of β, which under moderate assumptions and sufficient data is gaussian. Reporting results is typically in the form of probability intervals (credible sets and highest posterior density regions), quantiles of the posterior, and descriptions of probabilities of interest such as $P(\beta_i > 0)$.

The maximum likelihood estimate is equal to the Bayesian posterior mode with the appropriate uniform prior, and they are asymptotically equal given any proper (sums or integrates to one) prior. In many cases the choice of a prior is not especially important since as the sample size increases, the likelihood subsumes the prior. While the Bayesian assignment of a prior distribution for the unknown parameters can be seen as arbitrary, there are often strong arguments for particular forms of the prior: little or vague knowledge often justifies a diffuse or even uniform prior, certain probability models logically lead to particular forms of the prior, and the prior allows researchers to include additional information collected a priori.

5.3 Bayes Factor

Hypothesis testing can also be performed in the Bayesian setup. Suppose Θ_a and Θ_b represent two competing hypotheses about the state of some unknown parameter, β, which together form a partition of the sample space: $\Theta = \Theta_a \cup \Theta_b$; $\Theta_a \cap \Theta_b = \phi$. To begin, prior probabilities are assigned to each of the two outcomes: $\pi_a = P(\beta \in \Theta_a)$ and $\pi_b = P(\beta \in \Theta_b)$. This allows us to calculate the competing posterior distributions from the two priors and the likelihood function: $p_a = P(\beta \in \Theta_a|D,H_a)$ and $p_b = P(\beta \in \Theta_b|H_b)$. It is common to define the prior odds, π_a/π_b, and the posterior odds, p_a/p_b, as evidence for H_a versus H_b. A much more useful quantity, however, is $(\pi_a/\pi_b)/(p_a/p_b)$ which is called the *Bayes Factor*. The Bayes Factor is usually interpreted as odds favoring H_a versus H_b given the observed data. For this reason it leads naturally to the Bayesian analog of hypothesis testing between the two alternatives.

Classic references for Bayesian approaches include Box and Tao (1973), Press (1989), Jeffreys (1961), and Good (1950). More recent works on Bayesian models are Bernardo and Smith (1994), Pollard (1986), Lee (1989), and Jaynes 1996). Recent advances in Bayesian statistical computing, Markov Chain Monte Carlo (MCMC), have precipitated a dramatic increase in the application of Bayesian data analysis. Previously it was relatively easy to develop a model specification in which the resulting posterior distribution was either very difficult to obtain or completely intractable. However, new simulation techniques, primarily the Gibbs sampler and the Metropolis-Hastings algorithm, make it possible to numerically describe posterior densities from simulation evidence. These techniques, the most common forms of Markov Chain Monte Carlo (MCMC) procedures, as well as other statistical computing estimation techniques such as EM (Expectation-Maximization), Data Augmentation, and Monte Carlo methods in general, are revolutionizing the practice of statistics. For good book-length starting points on MCMC, see Gelman et al. (1995), Carlin and Lewis (1996), Gamerman (1997), Tanner (1996), and Gilks et al. (1996).

6 Substantively Weighted Analytical Techniques (SWAT)

The motivation of SWAT techniques is that not all public management cases are of equal interest to either scholars or practitioners. Practitioners might be interested in agencies that perform better than average given the constraints the agency faces, or in agencies that avoid failure in the face of complex tasks and uncooperative environments. The way to incorporate these practitioner concerns into useful research is use the discrepancy between model predicted performance and actual performance as an interesting measure unto itself.

In its most basic form, substantively weighted least squares(SWLS), SWAT uses a jack-knifed residual of 0.7 as a threshold of high performance (See Gill (1997) on generalizing this parameter). Rather than down weighting these extreme cases as robust regression analysis might do, in SWAT these cases are overweighted (or when investigating other subgroups down-weighted) to determine how these optimal performing agencies differ from the average agency. The SWLS form of SWAT is iterative in that one down-weights the average agencies in a series of regressions by increments of 0.1 until the average cases are counted as equal to only 0.1 high performing case. The changes found in these regressions should indicate the unique management elements that distinguish an excellent agency from a typical one.

The 0.7 threshold will generally designate about 20 percent of the cases as high performing with typical public administration data. The number of cases is a tradeoff. The fewer the number that are designated, the more the outcome will be the result of one or two cases which may or may not be generalizable. The analyst wants sufficient cases to be able to say that the relationships hold in a lot of agencies but not so many cases that we generalize to the mediocre cases.

From a statistical point of view, SWAT assumes that the regression coefficients vary across agencies. One of the differences between an excellent agency and a poor agency is that the excellent one gets far more output for a given level of input. That difference will show up in the weighted regressions when compared to the standard linear model regressions.

SWLS and other SWAT techniques do not estimate population parameters; that is, there is no longer a population to make inferences about. SWLS slopes should be thought of as indicators of how some agencies are different. The coefficients are qualitative indicators of roughly how much more (or less) the high performing bureaucracies get from their individual inputs. In this sense the resulting SWAT coefficients, which are displayed exactly like the standard linear model, are estimates of some hypothetical population of interest from which there are not enough easily identified cases.

The advantages of SWAT should not convince scholars to abandon ordinary least squares or regression diagnostics. These are obviously valuable research skills. Ordinary least squares and robust regression are the preferred technique to generalize from a sample to a population. They demonstrate how things are. SWLS or other SWAT techniques cannot be used to estimate relationships for a group of agencies; it is a technique used for performance isolation and recommendation. It is a qualitative technique that demonstrates how things might be.

Why is this approach useful? First consider the problem of defining high performing cases without a specific methodology (or what might be thought of as the “best practices” approach). Clearly highly advantaged cases benefit from the corresponding levels of explanatory variables. Therefore it is difficult to assert that a high performing case is doing well given a specific mix of levels without looking at the corresponding residual. Conversely, a highly

disadvantaged case may be performing extremely well relative to similarly affected cases but not relative to advantaged cases. In both scenarios, we are interested in residual outliers with all model specified explanations included.

Why is this approach better than segmenting out high performing cases and performing two analyses? Primarily because one needs to develop a model specification in order to get the residuals which determine high performance given relative benefits and hindrances. Those cases that most exploit their available resources are high performers, not simply those with high values of the dependent variable. Secondly, it is possible, even likely, that segmenting out these cases sufficiently reduces the sample size so that inference is difficult or unreliable. In the SWLS approach a focus is developed on the high performing cases where the others are reduced in emphasis to a "background." In a sense we get the primary information about the coefficient effects on the high performers cases where these results borrow strength from the full complement of cases.

The comprehensive guide to SWAT is Meier and Gill (2000). This book provides both a general introduction as well as coverage of some of the more complex theory and applications. SWAT has been used to investigate optimal performers (Meier and Keiser 1996), multiple goals (Meier, Wrinkle and Polinard 1999a), risk averse and failing organizations (Meier, Gill and Waller 2000), minority representation (Meier, Wrinkle and Polinard 1999b), and the differences between good agencies and exceptional ones (Gill and Meier 2000).

7 The Generalized Linear Model

The undeniable workhorse of public administration data analysis is the linear model (Meier and Brudney 1999). Researchers also employ an assortment of nonlinear regression tools such as logit and probit regression, event count models, truncated distribution models, and probability outcome models. These tools are imported singularly from other disciplines and treated as distinct topics. This has been a useful approach as the bulk of the empirical problems faced can be addressed with these methodologies. However, few in the profession are aware that these tools (and many more) are actually special cases of the *Generalized Linear Model*: a common method for producing model parameter estimates. So instead of having to find, understand, and apply a vast array of approaches with completely different terms and procedures, one can understand the single over-arching theory. Thus all of the particularistic tools can be

thought of more easily as special cases. Furthermore, because of its elegance and generality, it is inevitable that all of social sciences will adapt this framework as the dominant theoretical foundation of parametric models.

Treating various regression techniques completely separately, as nearly all texts in public administration data analysis do, leads to a very compartmented and necessarily limited view of the world. It also means that special procedures, specifications, and diagnostics must be learned separately. Conversely understanding a more encompassing theoretical basis through the generalized linear model allows a more universal and deeper theoretical view of empirical model building. Once the general framework is understood, then the optimum choice of model configuration is determined merely by the structure of the outcome variable and the nature of the dispersion.

7.1 The Exponential Family

In order to unify seemingly diverse specification forms, the generalized approach first recasts the chosen probability function into a consolidated exponential family form. This formalization is necessary to recharacterize familiar probability functions (discrete such as Poisson or binomial, and continuous such as gamma and gaussian) into a form that allows for a single theoretical treatment across seemingly disparate mathematical forms.

Suppose we consider a one-parameter conditional probability density function (continuous case, PDF) or probability mass function (discrete case, PMF) for the random variable Z_i of the form: $f(z_i|\zeta)$, read as "f of z sub-i given zeta". This presentation will focus for ease of discussion on the simplified case of a single parameter of interest, ζ, in the probability function, but this is certainly not a restriction of the generalized linear model. We index *z* by the subscript *i* to indicate that this is one random variable (the i^{th}) amongst a collection of many: we are rarely interested in single observations. This function, or more specifically this family of PDFs or PMFs, is classified as an exponential family type if it can be written in the form:

$$f(z_i|\zeta) = \exp[\underbrace{t(z_i)u(\zeta)}_{\text{interaction component}} + \underbrace{\log r(z_i) + \log s(\zeta)}_{\text{additive component}}]. \quad (10)$$

where: *r* and *t* are real-valued functions of z_i that do not depend on *i*, and s and *u* are real-valued functions of ζ that do not depend on z_i, and $r(z_i) > 0$; $s(\zeta) > 0 \ \forall \ z_i$. The first part of the right-hand side is labeled the "interaction component" because it is the component that reflects the product-indistinguishable relationship between functions of z_i and ζ. The second part of the right-hand side is labeled the "additive component" for obvious reasons. Despite this level of specificity, the form is very general because these restrictions are incredibly broad.

The form of (10) is specified for only one random variable, z_i. However, the exponential family form is preserved under random sampling meaning that the joint density function of an independent, identically distributed (i.i.d.) set of random variables, $\mathbf{Z} = \{z_1, z_2, \ldots, z_n\}$, is

$$f(\mathbf{z}|\zeta) = \exp\left[u(\zeta)\sum t(z_i) + \sum \log r(z_i) + n\log s(\zeta) \right] \quad (11)$$

where the summations range from 1 to n. So the joint distribution of a systematic random sample of variates with exponential family marginal distributions produces a joint an exponential family form.

The *canonical form* is a simplification that greatly facilitates estimation and inference. It is a transformation of the probability function that reduces the complexity of the symbolism and reveals important structure. If $t(z_i) = z_i$ in (11), then we say that this PDF or PMF is in a canonical form for the random variable Z. Otherwise we make the simple transformation: $y_i = t(z_i)$ to produce a canonical form. Furthermore, if $u(\zeta) = \zeta$ in (11), then we have a canonical form for ζ. If not, we can force a canonical form by transforming: $\theta = u(\zeta)$, and call θ the canonical parameter. The final form after these transformations is the following general expression, expressed as a joint distribution like (11):

$$f(\mathbf{y}|\theta) = \exp\left[\sum y_i\theta - n b(\theta) + \sum c(y_i) \right] \quad (12)$$

again where the summations range from 1 to n. The forms for $b(\theta)$ and

$c(y_i)$ are simply what results from the transformations on ζ and $\mathbf{z}$. They are intentionally left as general (i.e. the logs are not explicit), and the resulting form of $b(\theta)$ turns out to have great theoretical importance. In common cases we do not have to do the work of making transformations to achieve (12) as the canonical form is tabulated in many texts.

The form of θ in (12) is the "canonical link" between the original form and the θ parameterized form. The canonical link is used to generalize the linear model beyond the assumptions for normally distributed outcome variables. The forms of the canonical link and the normalizing constant for several common probability distributions are provided in Table 1. The term, $b(\theta)$ in (12) plays a key role in calculating the mean and variance of the distribution. These are more formally called the first two moments of the distribution. Table 1 also lists $b(\theta)$, the "normalizing constant" for the distributions.

Table 1: Canonical Links and Normalizing Constants

Distribution	Normalizing Constant, $b(\theta)$	Canonical Link, $\theta=g(\mu)$
Poisson	$\exp(\theta)$	$\log(\mu)$
Binomial	$n\log(1 + \exp(\theta))$	$\log(\mu/(1-\mu))$
Normal	$\theta^2/2$	μ
Gamma	$-\log(-\theta)$	$-1/\mu$
Negative Binomial	$r\log(1 - \exp(\theta))$	$\log(1-\mu)$
Inverse Gamma	$-(-2\theta)^{1/2}$	μ^{-2}

It can be easily seen from Table 1 that the standard linear model uses an identity canonical link function. In other words, in the simplest case when we "do nothing" to the generalized linear model, it reduces to the standard linear model, which is solved using ordinary least squares. So the familiar form of the linear model is nothing more than a special case of the generalized linear model.

7.2 Calculating the Mean and Variance of the Exponential Family

The generalization of the linear model is done by connecting the linear predictor, $\mathbf{X}\beta$, from a standard linear models analysis of the explanatory variables to the non-normal outcome variable through its mean function. Therefore the expected value plays a key theoretical role in the development of

generalized linear models. The expected value calculation of (12) with respect to the data (**Y**) is the wonderfully useful result: $EY = M/M\theta b(\theta)$. Furthermore, $VAR(Y) = M^2/M\theta^2 b(\theta)$ So all that is required from (12) to get the mean and variance of a particular exponential family of distributions is $b(\theta)$. This is an illustration of the value of expressing exponential family distributions in canonical form, since the derivatives of $b(\theta)$ immediately produce the mean and variance.

It is common to define a variance *function* for a given exponential family expression in which the θ notation is preserved for compatibility with the $b(\theta)$ form. The variance function is used in generalized linear models to indicate the dependence of the variance of Y on location and scale parameters. It is also important in developing useful residuals analysis. The variance function is simply defined as: $\tau^2 = M^2/M\theta^2 b(\theta)$, meaning that $VAR(Y) = M^2/M\theta^2 b(\theta)$ indexed by θ. Note that the dependence on $b(\theta)$ explicitly states that the variance function is conditional on the mean function.

7.3 The Generalization

Consider the standard linear model meeting the Gauss-Markov conditions (linear functional form, i.i.d. residuals with expected value zero and constant variance, and no correlation between any regressor and residual). This can be expressed as follows:

$$\underset{(nx1)}{\mathbf{V}} = \underset{(nxp)(px1)}{\mathbf{X}\beta} + \underset{(nx1)}{\varepsilon} \qquad (13)$$

$$\underset{(nx1)}{E(\mathbf{V})} = \underset{(nx1)}{\theta} = \underset{(nxp)(px1)}{\mathbf{X}\beta} \qquad (14)$$

The right-hand sides of the two equations contain: **X**, the matrix of observed data values, $\mathbf{X}\beta$, the "linear structure vector", and ε, the error terms. The left-hand side contains: $E(\mathbf{V}) = \theta$, the vector of means: the systematic component. The variable, **V**, is distributed i.i.d. normal with mean θ, and constant variance σ^2. Now suppose we generalize slightly this well known form with a new "linear predictor" based on the mean of the outcome variable, **Y**, which is no longer required to be normally distributed or even continuous:

$$g(\mu) = \theta = \mathbf{X}\beta \qquad (15)$$

(nx1) (nx1) (nxp)(px1)

Here g() is required to be an invertible, smooth function of the mean vector μ of **Y**. Information from the explanatory variables is now expressed in the model only through the link from the linear structure, $\mathbf{X}\beta$, to the linear predictor, $\theta = g(\mu)$, controlled by the form of the link function, g(). This link function connects the linear predictor to the mean of the outcome variable not directly to the expression of the outcome variable itself, so the outcome variable can now take on a variety of non-normal forms. The link function connects the stochastic component which describes some response variable from a wide variety of forms to all of the standard normal theory supporting the systematic component through the mean function: $g^{-1}(g(\mu)) = g^{-1}(\theta) = g^{-1}(\mathbf{X}\beta) = \mu = E(\mathbf{Y})$. So the inverse of the link function ensures that $\mathbf{X}\beta$ maintains the Gauss-Markov assumptions for linear models and all of the standard theory applies even though the outcome variable no longer meets the required assumptions.

The generalization of the linear model now has four components derived from the expressions above.

I. **Stochastic Component**: **Y** is the random or stochastic component which remains distributed i.i.d. according to a specific exponential family distribution with mean μ.

II. **Systematic Component**: $\theta = \mathbf{X}\beta$ is the systematic component with an associated Gauss-Markov normal basis.

III. **Link Function**: the stochastic component and the systematic component are linked by a function of θ which is exactly the canonical *link function*, summarized in Table 1. We can think of $g(\mu)$ as "tricking" the linear model into thinking that it is still acting upon normally distributed outcome variables.

IV. **Residuals**: Although the residuals can be expressed in the same manner as in the standard linear model, observed outcome variable value minus predicted outcome variable value, a more useful quantity is the deviance residual described in detail below.

7.4 Estimation: Iterative Weighted Least Squares

Even though it is very common for the variance structure to be dependent on the mean function, it is relatively rare to know the exact form of the dependence to estimate the covariances. A solution to this problem is to iteratively estimate regression weights, improving the estimate on each cycle using the mean function. Since $\mu = g^{-1}(\mathbf{X}\beta)$, then the coefficient estimate, $\hat{\beta}$, provides a mean estimate and vice versa. The iterative weighted least squares algorithm alternately estimates these quantities using progressively improving weights. Under very general conditions, satisfied by the exponential family of distributions, iterative weighted least squares finds the mode of the likelihood function, thus producing the maximum likelihood estimate of the unknown coefficient vector, $\hat{\beta}$. For a detailed explanation of the procedure and its theoretical justification, the reader is directed to del Pino (1989), Gill (2000), and Green (1984).

7.5 The Deviance Function and Deviance Residuals

By far the most useful category of residuals for the generalized linear model is the deviance residual. This is also the most general form. A common way to look at model specification is the analysis of the likelihood ratio statistic comparing a proposed model specification relative to the saturated model (n data points, n specified parameters, using the exact same data and link function). The difference in fit is generally called the summed deviance. Since this deviance is composed of the contributions from each data point and the difference between summarizing with a relatively small subset of parameters and one parameter for every data point, then these individual deviances are directly analogous to residuals.

Starting with the log likelihood for a proposed model, add the "^" notation as a reminder that it is evaluated at the maximum likelihood values:

$L(\hat{\theta}|\mathbf{y}) = \sum \left[y_i \hat{\theta} \; b(\hat{\theta}) \right] + c^*(y_i)$. Now also consider the same log likelihood function with the same data and the same link function, except that it now has n coefficients for the n data points, i.e. the saturated model log likelihood function with the "~" function to denote the n-length θ vector:

$L(\tilde{\theta}|\mathbf{y}) = \sum \left[y_i \tilde{\theta}\ b(\tilde{\theta}) \right] + c^*(y_i)$. This is the highest possible value for the log likelihood function achievable with the given data, **y**. Yet it is also generally useless analytically except as a benchmark. The deviance function is defined as minus twice the log likelihood ratio (i.e. the arithmetic difference, since both terms are already written on the log metric):

$$D(\theta,\mathbf{y}) = \sum \left[L(\tilde{\theta}|\mathbf{y} - L(\hat{\theta}|\mathbf{y}) \right] = \sum \left[y_i(\tilde{\theta} - \hat{\theta}) - (b(\tilde{\theta}) - b(\hat{\theta})) \right]. \quad (16)$$

This is a measure of the summed difference of the data-weighted maximum likelihood estimates and the $b(\theta)$ parameters. Thus the deviation function gives a measure of the trade-off between a saturated model which fits every single data point, assigning all variation to the systematic component, and a proposed model which reflects the researcher's belief about the identification of the systematic and random components. Hypothesis tests of fit are performed using the asymptotic property that $D(\theta,\mathbf{y}) \sim \chi^2_{n-p}$ (although the asymptotic rate of convergence varies dramatically depending on the exponential family form). However, for dichotomous and count outcome data, the convergence of the deviance function to a χ^2_{n-p} is relatively slow. In cases involving such outcome variables, it is strongly advised to add or subtract 1/2 to each outcome variable in the direction of the outcome variable mean. This continuity correction greatly improves the distributional result (Pierce and Schafer 1971, 1986). Observe once again that the $b(\theta)$ function plays a critical role.

Although calculating $D(\theta,\mathbf{y})$ is relatively straightforward, we usually do not need to do this as many texts provide the result for frequently used PDFs and PMFs (Gill 2000, Jϕrgensen 1997, McCullagh and Nelder 1989). A utility of the deviance function is that it also allows a look at the individual deviance contributions in an analogous way to linear model residuals. The single point deviance function is just the deviance function for the yth point (i.e. without the summation): $d(\theta,y_i) = \left[y_i(\tilde{\theta} - \hat{\theta}) - (b(\tilde{\theta}) - b(\hat{\theta})) \right]$. To define the deviance residual at the y_i point, we take the square root:

$$R_{Deviance} = \frac{(y_i - \mu_i)}{|y_i - \mu_i|}(|d(\theta, y_i)|)^{1/2} \quad (17)$$

where $\frac{(y_i - \mu_i)}{|y_i - \mu_i|}$ is just a sign-preserving function.

This section has very briefly summarized the theory and major components of the generalized linear model. It should be apparent that it is both a broad construct, encompassing most parametric forms of interest to public administration data analysts, and a complete system of analysis that includes a specification process, a computational algorithm, and a criteria for analyzing fit. It is not implied that this discussion of generalized linear models is anywhere near complete, instead it is a short preface to a large body of applied and theoretical literature. The classic book-length work is McCullagh and Nelder (1989). Two very accessible works are Dobson (1990) and Lindsey (1997). An excellent, but more complex, book is that of Fahrmeir and Tutz (1994). An exposition focused on social science data analysis and the theory of generalized linear models is Gill (2000).

8 Incentives to Develop Methods

Any methodological manifesto in public administration should address the problem of implementation. Prior calls to battle have been issued but relatively little has changed. For progress to be made in public administration developing its own methods, the current incentives of the profession need to be altered so that individuals will invest the time necessary to develop new methods or alter existing methods to the specific needs of public administration.

8.1 A Publication Outlet

Any scholar investing time in developing new methods or translating methods into public administration needs a publication outlet that is receptive to such an endeavor and is likely to be read by others in public administration. Our perception is that most in public administration do not read the *Journal of the American Statistical Association, Econometrica* or even the workshop section of the *American Journal of Political Science*. If one did religiously read such publications, the valuable information gained would be swamped by the volume of dross; much of what appears in those pages is unlikely to have any useful applications in public administration.

Without a new outlet for methods-related work, public administration scholars working in this area face the problem that they must make a substantive contribution as well as illustrate a new method, a not enviable task made difficult by editorial demands to dumb down manuscripts for a practitioner readership. A separate journal devoted to methods in public administration is not likely to be feasible. A better solution would be to have workshop sections, patterned after that in the *American Journal of Political Science*, in one or more public administration journals where individuals could present and illustrate methods.

9 Final Comments

9.1 Winning By Losing

An interesting public policy scenario developed during the late 1980s into the 1990s with regard to high definition television (HDTV). The three primary industrialized sectors, Europe, Japan, and the U.S., were anxious to produce a working system, and none of the key governmental and industrial players in the three centers were able to agree on a global technical standard. Key decisions needed to be made such as the frequency arrangement and whether or not the system would be analog or digital. The Japanese were more aggressive and had a technical lead. In fact they developed a working analog system in Japan, albeit an expensive and exclusive one. Conversely, the United States was mired in inter-agency conflict and political battles between broadcasters and electronic manufacturers. As a result the Japanese had a working system when there was none in Europe and the U.S. The irony of this story is that the U.S. actually won this competitive economic battle because the right type of system was the superior digital standard rather than the analog. So by being behind, politically and technically, the U.S. avoided making a significant financial investment in an inferior earlier standard, and eventually set the standard for the world system, which advantaged U.S. corporations and customers. We see the current state of public administration methodology in this light. Clearly the field is behind related fields such as economics, political science, psychology, and sociology. We should, however, turn this into an advantage rather than a deficit by avoiding some of the unproductive paths that these other fields have taken (null hypothesis significance testing, non-integrated approaches, content analysis, excessive scaling, case study generalizations, confusion about causality, etc.), and pursue the areas that have been productive or promising (times series analysis, SWAT, Bayesian methods, GLM).

9.2 The Empirical Proletariat

Public administration as an empirical research field is dominated by case studies and description in general. Evidence to support this statement is easily found by browsing the field's journals at random or looking at the ASPA annual meeting program. One sees a gamut of particularistic work ranging in importance from "Improving Garbage Collection in Topeka" to "Organizing the Executive Office of the President" (actual paraphrased titles). As a result a large proportion of the high quality quantitative work addressing questions in public administration, and public management in particular, are published in political science, sociology, or business administration journals. This clearly damages the integrity of the boundaries, and the legitimacy, of public administration as a distinct academic field. Thus part of the message of this manifesto is a call for more rigorous mathematical, statistical, and formal theoretic applications to questions in public administration *to be published in public administration*. In another perhaps more famous manifesto, Marx and Engels state: "All previous historical movements were movements of minorities, or in the interest of minorities." The approaches that we advocate clearly make us a minority; we would now like to create a movement.

9.3 The Methodological Bourgeoisie

Heinz Eulau once described research in public administration as "an intellectual wasteland suffering from undue constriction of scope, theory, and method" (Bobrow et al. 1977). This less than tactful statement was made well over twenty years ago, and we believe it is currently only one third correct. It is rare to the point of surprise that a new methodological contribution in the social sciences comes from the public administration literature. This is not to say that important, creative, and occasionally sophisticated methods are not employed in public administration research. Instead we mean that there are few instances where a methodology developed to address a research question in public administration is subsequently applied in other fields. However, as per the HDTV analogy above, there is great opportunity accelerate the production of research tools. Quoting again from that other manifesto, "The bourgeoisie cannot exist without constantly revolutionizing the instruments of production."

9.4 Free the Bound Periodicals

We disdain to conceal our views and aims with regard to publishing outlets in public administration. The current mean methodological level of journals in the field is significantly lower than every other social science. The reasons for

this are varied but the core issue is that many journals pander to the lower levels of research sophistication held by public administration practitioners. The so-called "flagship" journal of the field is the most flagrant offender in this regard. It is not that the interests and concerns held by practitioners are necessarily bad for the field, rather that practitioners are holding back the methodological and theoretical discussion in the premier journals.

Why are we picking on practicing public administrators? Well actually we are not. But by binding together publishing outlets that affect academic careers with the more nuts-and-bolts interests of those who manage programs, balance budgets, and implement policies, we are making a deliberate normative statement about the orientation of the discipline. This orientation means that methodologically sophisticated public administration research will end up appearing in the *American Journal of Political Science*, the *American Sociological Review*, or *Administrative Science Quarterly* before any public administration journal. Actually our reasoning is somewhat circular as some researchers that feel constrained by the orientation of the journals are, to a great extent, the ones who teach these practitioners through MPA programs. So perhaps fault lies with the NAASPA induced curricular rigidity that makes it difficult to teach quantitative approaches beyond tabular analysis and the linear model.

As the title of this paper implies, we have strong opinions about the direction of the discipline. Public administration is out of balance relative to other social sciences, and in general quite far behind with regard to analyzing data in meaningful ways. This is certainly a correctable situation, and we have provided a number of paths to get there. We encourage a greatly enhanced focus on empiricism and rigorous quantitative approaches. "The proletarians have nothing to lose but their chains."

10 References

Amemiya, Takeshi. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Barnett, Vic. 1973. *Comparative Statistical Inference*. New York: John Wiley & Sons.

Berger, James O., B. Boukai, and Y. Wang. 1997. Unified Frequentist and Bayesian Testing of a Precise Hypothesis." *Statistical Science* 12, 133-60.

Bernardo, J. M., and A. F. M. Smith. 1994. *Bayesian Theory*. New York: John Wiley & Sons.

Berry, Willam D., Richard C. Fording and Russell L. Hanson. 1997. Reassessing the "Race to the Bottom Thesis: A Spatial Dependence Model of State Welfare Policy" Paper delivered at the 1997 Annual Meeting of the American Political Science Association, August 28-31, 1997.

Bobrow, Davis B., Heinz Eulau, Martin Landau, Charles O. Jones, and Robert Axelrod. 1977. "The Place of Policy Analysis in Political Science: Five Perspectives." *American Journal of Political Science* 31 (May), 419-23.

Box, G. E. P., and G. C. Tao. 1973. *Bayesian Inference in Statistical Analysis*. New York: Wiley Classics.

Brenner-Golomb, Nancy. 1993. "R. A. Fisher's Philosophical Approach to Inductive Inference." In G. Kerenand C. Lewis, eds. *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carlin, Bradley P., and Thomas A. Louis. 1996. *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman & Hall.

Casella, George, and Roger L. Berger. 1990. *Statistical Inference*. Belmont, CA: Wadsworth & Brooks/Cole.

Cohen, Jacob. 1994. "The Earth is Round ($p < .05$)." *American Psychologist* December, 12, 997-1003.

Cohen, Jacob. 1962. "The Statistical Power of Abnormal-Social Psychological Research: A Review." *Journal of Abnormal and Social Psychology* 65, 145-53.

Davidson, Russell and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

De Boef, Suzanna, and Jim Granato. 1999. "Testing for Cointegrating Relationships With Near-integrated Data." *Political Analysis* 8 (No.1), 99-117.

del Pino, Guido. 1989. "The Unifying Role of Iterative Generalized Least Squares in Statistical Algorithms." *Statistical Science* 4: 394-408.

Dobson, Annette J. 1990. *An Introduction to Generalized Linear Models*. New York: Chapman & Hall.

Durant, Robert F., Jerome S. Legge, and Antony Moussios. 1998. "People. Profits, and Service Delivery: Lessons from the Privatization of British Telecom." *American Journal of Political Science* 42 (January): 117-140.

Durr, Robert H. 1993. "What Moves Policy Sentiment?" *American Political Science Review* 87 (March): 158-70.

Fahrmeir, Ludwig, and Gerhard Tutz. 1994. *Multivariate Statistical Modeling Based on Generalized Linear Models*. New York: Springer.

Falk, R., and C. W. Greenbaum. 1995. "Significance Tests Die Hard." *Theory and Psychology* 5, 396-400.

Fisher, Sir Ronald A. 1925. "Theory of Statistical Estimation." *Proceedings of the Cambridge Philosophical Society* 22: 700-25.

Freeman, John. 1983. "Granger Causality and Time Series Analysis of Political Relationships." *American Journal of Political Science* 27: 327-58.

Gamerman, Dani. 1997. *Markov Chain Monte Carlo*. New York: Chapman & Hall.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. New York: Chapman and Hall.

Gigerenzer, Gerd. 1987. "Probabilistic Thinking and the Fight Against Subjectivity." In Krüger, Lorenz, Gerd Gigerenzer, and Mary Morgan, eds. *The Probabilistic Revolution. Volume 2*. Cambridge, MA:MIT.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. New York: Chapman and Hall. Gill,

Gill, Jeff,. 2000. *Generalized Linear Models: A Unified Approach*. Newbury Park, CA: Sage, QASS Series.

Gill, Jeff,. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52: 647-74.

Gill, Jeff,. 1997. "Generalized Substantively Reweighted Least Squares Regression" Political Methodology Working Paper archive, http://www.polmeth.calpoly.edu

Gill, Jeff, and Kenneth J. Meier. "Ralph's Pretty Good Grocery Versus Ralph's Super Market: Separating Excellent Agencies from the Good Ones." *Public Administration Review* (forthcoming).

Good, I. J. 1950. *Probability and the Weighing of Evidence*. New York: Hafner.

Granger, Clive W. J. 1969. "Investigating Causal Relationships by Econometric Methods." *Econometrica* 37: 424-38.

Granger, Clive W. J., P. Newbold. 1976. "Forecasting Transformed Series." *Journal of the Royal Statistical Society* Series B (38, No.2), 189-203.

Green, P. J. 1984. "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives." *Journal of the Royal Statistical Society*, Series B 46: 149-92.

Greene, William. 1999. *Econometric Analysis*. Fourth Edition. New York: Prentice Hall.

Greenwald, Anthony G. 1975. "Consequences of Prejudice Against the Null Hypothesis." *Psychological Bulletin* 82, 1-20.

Gujarati, Damodar N. 1995. *Basic Econometrics*. New York: McGraw Hill.

Hedge, David, Donald Menzel, and George H. Williams. 1988. "Regulatory Attitudes and Behavior." Western Political Quarterly 41 (June), 323-40.

Hogg, Robert V., and Allen T. Craig. 1978. *Introduction to Mathematical Statistics*. New York: Macmillan.

Jaynes, E. T. 1996. *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.

Jeffreys, Harold. 1961. *The Theory of Probability*. Oxford: Clarendon Press.

Jϕrgensen, Bent. 1997. *The Theory of Dispersion Models*. New York: Chapman & Hall.

Judge, G., C. Hill, W. Griffiths, T. Lee, H. Lutkepol. 1982. *An Introduction to the Theory and Practice of Econometrics*. New York: John Wiley & Sons.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30 (3) 666-87.

Krause, George A. 1996. "The Institutional Dynamics of Policy Administration: Bureaucratic Influence Over Securities Regulation." *American Journal of Political Science* 40 (November): 1083-1121.

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

Lee, Peter M. 1989. *Bayesian Statistics*. Oxford: Oxford University Press.

Lehmann, E. L. 1986. *Testing Statistical Hypotheses*. Second Edition. New York: Springer.

Lewis, Gregory B., and David Nice. 1994. "Race, Sex and Occupational Segregation in State and Local Governments." *American Review of Public Administration* 24, 393-410.

Li, Hongyi, and Maddala, G. S. 1996. "Bootstrapping Time Series Models." *Econometric Reviews* 15,115-58.

Lindsey, James K. 1997. Applying Generalized Linear Models. New York: Springer-Verlag. Lindsay, R. M. 1995. "Reconsidering the Status of Tests of Significance: An Alternative Criterion of Adequacy." *Accounting, Organizations and Society* 20, 35-53.

Lutkepohl, Helmut. 1982. "Non-Causality Due to Omitted Variables." *Journal of Econometrics* 19: 367-78.

Maddala, G. S. 1992. *Introduction to Econometrics*. Second Edition. New York: Macmillan.

Macdonald, Ranald R. 1997. "On Statistical Testing in Psychology." *British Journal of Psychology* 88, No.2 (May), 333-49.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. Second Edition. New York: Chapman & Hall.

Meier, Kenneth J., and Jeff, Gill. 2000. *Substantively Weighted Analytical Techniques*. Boulder, CO: Westview Press.

Meier, Kenneth J., Jeff Gill and George Waller. 2000. "Optimal Performance versus Risk Aversion: An Application of Substantive Weighted Least Squares." In

Hal Rainey, Jeffrey L. Brudney and Laurence O'Toole, *Advancing Public Management: New Developments in Theory, Methods and Practice*. Washington, DC: Georgetown University Press, 2000.

Meier, Kenneth J., and Jeffrey L. Brudney. 1999. *Applied Statistics for Public Administration*. New York: Harcourt Brace.

Meier, Kenneth J., and Lael R. Keiser. 1996. "Public Administration as a Science of the Artificial: A Methodology for Prescription." *Public Administration Review* 56: 459-66.

Meier, Kenneth J. and Kevin B. Smith. 1994. "Representative Democracy and Representative Bureaucracy: Examining the Top-Down and the Bottom-Up Linkages." *Social Science Quarterly* 75 (December): 790-803.

Meier, Kenneth J., J.L. Polinard, and Robert D. Wrinkle. 1999. "Politics, Bureaucracy and Farm Credit." Public Administration Review 59 (July/August), 293-302.

Meier, Kenneth J., Robert D. Wrinkle, and J.L. Polinard. 1999a. "Equity and Excellence in Education: A Substantively Reweighted Least Squares Analysis." *American Review of Public Administration* 29 (March): 5-18.

Meier, Kenneth J., Robert D. Wrinkle, and J.L. Polinard. 1999b "Representative Bureaucracy and Distributional Equity: Addressing the Hard Question." *Journal of Politics* 61 (November 1999).

Miller, Alan J. 1990. *Subset Selection in Regression*. New York: Chapman & Hall.

Miller, Will, Brinck Kerr, and Margaret Reid. 1999. ``A National Study of Gender-Based Occupational Segregation in Municipal Bureaucracies: Persistence of Glass Walls? *Public Administration Review* 59 (3), May/June, 218-29.

Neyman, Jerzy, and Egon S. Pearson. 1933a. "On the Problem of the Most Efficient Test of Statistical Hypotheses." Philosophical Transactions of the Royal Statistical Society A 231, 289-337.

Neyman, Jerzy, and Egon S. Pearson. 1933b. "The Testing of Statistical Hypotheses in Relation to Probabilities a priori". Proceedings of the Cambridge Philosophical Society 24, 492-510.

Neyman, Jerzy, and Egon S. Pearson. 1936. "Contributions to the Theory of Testing Statistical Hypotheses." Statistical Research Memorandum 1, 1-37.

Newman, Meredith Ann. 1993. "Career Advancement: Does Gender Make a Difference." *American Review of Public Administration* 53, 361-84.

Newton, Isaac. 1728. *The Chronology of Ancient Kingdoms Amended*. London: Tonson, Osborn, and Longman.

Oakes, M. 1986. *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: John Wiley & Sons.

Peers, H. W. 1971. "Likelihood Ratio and Associated Test Criteria." *Biometrika* 58: 577-89.

Perry, James L., and Kenneth L. Kraemer. 1986. "Research Methodology in the Public Administration Review, 1975-1984." *Public Administration Review* 46(May/June) (1986): 215-226.

Peterson, Paul E., and Mark Rom. 1989. "American Federalism, Welfare Policy, and Residential Choices." *American Political Science Review* 83 (No.3), 711-728.

Pierce, Donald A., and Daniel W. Schafer. 1986. "Residuals in Generalized Linear Models." *Journal of the American Statistical Society* 81: 977-86.

Pindyck, Robert s. and Daniel L. Rubinfeld. 1991. *Econometric Models and Econometric Forecasts*. Third Edition. New York: McGraw Hill.

Pollard, P., and J. T. E. Richardson. 1987. "On the Probability of Making Type One Errors." *Psychological Bulletin* 102, (July), 159-63.

Pollard, W. E. 1986. *Bayesian Statistics for Evaluation Research*. Newbury Park, CA: Sage.

Popper, Karl. 1968. *The Logic of Scientific Discovery*. New York: Harper and Row.

Press, S. J. 1989. Bayesian Statistics: Principles, Models, and Applications. New York: John Wiley & Sons.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." In Peter V. Marsden ed. *Sociological Methodology*. Cambridge, MA: Blackwell.

Rohatgi, V. K. 1976. *An Introduction to Probability Theory and Mathematical Statistics*. New York: John Wiley & Sons.

Rosenthal, Robert. 1979. "The 'File Drawer Problem' and Tolerance for Null Results." *Psychological Bulletin* 86: 638-41.

Rosenthal, R. and J. Gaito. 1963. "The Interpretation of Levels of Significance by Psychological Researchers." *Journal of Psychology* 55: 33-9.

Rosenthal, Robert and Donald Rubin. 1988. "Comment: Assumptions and Procedures in the File Drawer Problem." *Statistical Science* 3, 120-5.

Rosenthal, Robert and Donald Rubin. 1978. "Interpersonal Expectancy Effects: The First 345 Studies." *The Behavioral and Brain Sciences* 3: 377-86.

Rozeboom, William W. 1960. "The Fallacy of the Null Hypothesis Significance Test." *Psychological Bulletin* 57, 416-28.

Schmidt, Frank L. 1996. "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers." *Psychological Methods* 1, 115-129.

Schmidt, Peter. 1976. *Econometrics*. New York: Marcel Dekker.

Smith, Kevin B. 1997. "Explaining Variation in State-Level Homicide Rates: Does Crime Policy Pay?" *Journal of Politics* 59: 350-67.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cam-bridge, MA: Harvard University Press.

Stigler, Stephen M. 1977. "Eight Centuries of Sampling Inspection: The Trial of the Pyx." *Journal of the American Statistical Association* September, 72, 493-500.

Stimson, James. 1985. "Regression in Space and Time: A Statistical Essay." *American Journal of Political Science* 29: 914-47.

Stuart, Alan, and Keith Ord. 1994. *Kendall's Advanced Theory of Statistics, Volume I: Distribution Theory*. Sixth Edition. London: Edward Arnold.

Tanner, Martin A. 1996*. Tools for Statistic Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. New York: Springer.

Thurman, W. N. and M. E. Fisher. 1988. "Chickens, Eggs, and Causality, or Which Came First." *American Journal of Agricultural Economics* 70: 237-48.

Tufte, Edward R. 1977. "Political Statistics for the United States: Observations on Some Major Data Sources." *American Journal of Political Science* 21 (March), 305-14.

Wilkerson, Matt, and Mary R. Olson. 1997. "Misconceptions About Sample Size, Statistical Significance, and Treatment Effect." *Journal of Psychology* forthcoming.

Wood, B. Dan. 1992. "Modeling Federal Implementation as a System: The Clean Air Case." *American Journal of Political Science* 36 (February): 40-67.