Journal of Public Administration, 28: 5–35, 2005

ISSN 0190-0692 print / 1532-4265 online
DOI: 10.1081/PAD-200044556

Routledge
Taylor & Francis Group

Bayesian Inference in Public Administration Research: Substantive Differences from Somewhat Different Assumptions

Kevin Wagner*
Department of Political Science, University of Florida, Gainesville, Florida, USA 32611

Jeff Gill
Department of Political Science, University of California-Davis, Davis, California, USA 95616-8682

Abstract: The purpose of this article is to point out that the standard statistical inference procedure in public administration is defective and should be replaced. The standard classicist approach to producing and reporting empirical findings is not appropriate for the type of data we use and does not report results in a useful manner for researchers and practitioners. The Bayesian inferential process is better suited for structuring scientific research into administrative questions due to overt assumptions, flexible parametric forms, systematic inclusion of prior knowledge, and rigorous sensitivity analysis. We begin with a theoretical discussion of inference procedures and Bayesian methods, then provide an empirical example from a recently published, well-known public administration work on education public policy.

STATISTICAL INFERENCE PUBLIC ADMINISTRATION: A REPORT CARD

Methodology in public administration stands at a crossroads. The choice is to continue to employ a dated and inappropriate device for determining the reliability of statistical findings or to reevaluate and seek better suited inferential tools. In this article we will demonstrate that the dominant Null Hypothesis Significance Test (NHST) is broken and cannot be repaired, and then we will argue that the appropriate direction for our field is down the Bayesian road.

Correspondence: Kevin Wagner, Department of Political Science, University of Florida, 234 Anderson Hall, Gainesville, Florida, 32611, USA; E-mail: kwagner@polisci.ufl.edu

It is not a secret that empirical research in public administration has lagged behind related fields.[1] The mean methodological sophistication in public administration journals is far below those of sociology, political science, anthropology, and economics. However, this does not have to be. We do not prescribe here to Heinz Eulau's assessment of public administration research as "an intellectual wasteland."[2] In fact, we believe quite the opposite: public administration research provides more practical and germane prescriptive findings than all of these other listed fields. What we need, however, is freedom from the flawed but pervasive inferential procedure for making statistical inference in public administration research.

Borrowing Our Neighbor's Tools

Public administration relies heavily on other related disciplines for its methodological tools, despite the *specific* research challenges in our field. This has several ramifications. These other disciplines, primarily political science and sociology, are driven by fundamentally different questions. In particular, there is much less interest in analyzing survey research in public administration. This is important methodologically because the data sets analyzed more commonly in public administration are complete population data rather than samples. It has been shown that treating population data like sample data with regards to social science analysis leads to a number of negative pathologies.[3]

Data generated as a population rather than from a repeatable known probability process can be described as non-stochastic.[4] This is inherently problematic because traditional statistical methods are predicated on the assumption that the data are created by a repeatable mechanism. In a standard frequentist model, a sample for statistical purposes must be a single manifestation from many possible outcomes drawn from an unchanging probability distribution.[5] In practice, this is accomplished through random or probability sampling. An inference about a particular parameter is dependent upon this process and the resulting sampling distribution generated by a mechanism for drawing the sample that can be repeated multiple times. Where the dataset is the complete population, standard sample-derived inferences are not simply difficult, they are inapplicable as they are founded on inherently untrue population assumptions.[6]

On the other side of the spectrum, economics offers public administration even less, methodologically. There has lately been an increasing emphasis on theory in economics at the expense of applied empirical work. The result of this movement is that models in that field rely increasingly on strong distributional assumptions and abstract theoretical specifications. Consequently, issues of measurement error, robustness, specification sensitivity, and

graphical analysis are downplayed.[7] These are issues of particular importance in public administration research where scholars often seek to make prescriptive findings.

Public administration is arguably unique (omitting economics) among associated disciplines in having this strong prescriptive orientation. This results in a problem associated with wholesale importation of social science methodology. While political scientists generally are very happy to further the understanding of previous elections or governments, and sociologists may be quite pleased in explaining past social movements and behaviors, public administration often seeks to inform practitioners and interested scholars about how managerial decisions *should* be made. Methodologically, we know that there is greater stochastic uncertainty involved in making predictions rather than simply fitting data, thus making empirical research in public administration more challenging than expected.

In addition, public administration research often suffers from issues related to collinearity, since scholars regularly obtain datasets with a large number of variables that can be causally related. In the case of education research, teacher salary, class size or even poverty level in the district can reflect similar or related information.[8] This can create difficulties in a linear model, as collinear explanatory variables carry little independent information, and the least squares estimator does not then provide a means to distinguish information concerning one coefficient from another.[9]

Finally, it is important to note that a huge proportion of published research in public administration is case-study oriented. There is nothing wrong with this approach, and it has furthered our knowledge on a number of fronts. However, this focus has impeded the field's pursuit of underlying causal phenomena, which is often pursued with formal or statistical analysis. Further, the dominant inferential methods provide no systematic means to incorporate into new work knowledge gained from such case-study research. Clearly, public administration would benefit form a balanced approach to research design and theory testing.

The Completely Bankrupt Process of Null Hypothesis Significance Testing

The dominant approach to hypothesis testing in public administration (and most social sciences) does not work. We are not saying it is not optimal, it could be improved, some things about it need to be fixed, or that many researchers just do not apply it appropriately. Instead, we will now show that *it is wrong and it needs to be replaced.*

The ubiquitous NHST (Null Hypothesis Significance Testing) is an unintentional synthesis of the Fisher *test of significance* and the Neyman-Pearson *hypothesis test*. The procedure works as follows:

1. Two hypotheses are posited: a null or restricted hypothesis (H_0), which competes with an alternative or research hypothesis (H_1), each describing complementary notions about some social or administrative phenomenon.
2. The hypotheses are operationalized through statements about some parameter, β. Often (say, in regression models) these are statements such as:

$$H_0 : \beta = 0 \quad H_1 : \beta \neq 0. \tag{1.1}$$

3. 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.
4. 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.
5. 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.
6. 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.

This process is a synthesis of two important but incompatible procedures in modern statistics. Fisher produces significance levels *from the data* whereas Neyman and Pearson posit a test-oriented decision process that confirms or rejects hypotheses at *a priori* specified levels. However, the NHST tries to blend these two disparate approaches, leading to improper conclusions about the data. In Fisher hypothesis testing, no explicit complementary hypothesis to H_0 is identified. For him, the p-value that results from the model and the data is evaluated alone as the strength of the evidence for the research hypothesis. There is no fixed deciding line. There is also no notion of the *power of the test* in Fisher's procedure: the probability of correctly rejecting H_0 is very important in the Neyman-Pearson procedure. In addition, there is no overt decision in favor of H_0 in the NHST (instructors in introductory social science statistics courses always admonish students never to *accept* the null. Conversely, Neyman-Pearson tests identify two co-equal complementary hypotheses: Θ_A and Θ_B where rejection of one implies immediate acceptance of the other and this rejection is based on a predetermined α level. Therefore, one of two decisions must always be accepted.

The current paradigm in the social sciences blends these two approaches by pretending to select α *a priori*, but actually uses ranges of p-values to evaluate

strength of evidence. This allows researchers to include the alternate hypothesis without having to search for more powerful tests (often a difficult task). The test also 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. Sometimes the null hypothesis is presented only as a null relationship: $\beta = 0$ (i.e., no effect), but Fisher really intended the null hypothesis simply as something to be nullified. Conflictingly, 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. This later assertion is incorrect in this context since the probability is predicated on the null being true, and because only one sample is involved, there is no long-run probability achieved.

There are several misconceptions that result from using the null hypothesis significance test. First, many incorrectly believe that the smaller the p-value, 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, but the NHST first posits H_0 as true then asks what is the probability of observing these or more extreme data. Second, it is common to confuse the decision process with the strength of evidence: the NHST interpretation of hypothesis testing does not distinguish between 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."[10] Third, the infinite number of alternatives to the null are not considered in any one test, thus not being ruled out as possible explainers of phenomenon. So, failure to find evidence for the one research hypothesis does not rule out any others, and therefore does not support the null. Fourth, the core underlying logic is flawed. The basic strategy is to make an assumption, observe some real-world event, and then check the consistency of the assumption given this observation. But the order of conditionality is exactly the reverse, and we know from Bayes Law that these cannot be equal.

Thus, the null hypothesis simply does not provide what researchers and consumers of research in public administration want it to provide. We would like to make probabilistic prescriptions about various phenomena and possible courses of action that administrators might take. What is needed is a more intuitive way to interpret data and models. In the next section we introduce the Bayesian statistical paradigm and argue that this is the best vehicle for these purposes.

THE BAYESIAN WAY

The Bayesian process of data analysis allows researchers to incorporate systematically previous knowledge into a statistical model and to make *probability* statements concerning the outcome. More directly, Bayesian methodology is characterized by three primary attributes: a willingness to assign prior

distributions to unknown parameters, the use of Bayes rule to obtain a posterior distribution for unknown parameters and missing data conditioned on observable data, and the description of inferences in probabilistic terms. The core philosophical foundation of Bayesian inference is the consideration of both observables and parameters as random quantities. A primary advantage of this approach is that there is no restriction to building complex models with multiple levels and many unknown parameters. Because model assumptions are much more conspicuous in the Bayesian setup, readers can more accurately assess model quality and specifications.

Bayesian inference differs from standard methods in that it is based on fundamentally different assumptions about collected data and unknown parameters.[11] In the Bayesian view, quantities are divided into two groups: observed and unobserved. Observed quantities consist of the data and known constants. Unobserved quantities consist of parameters of interest to be estimated, missing data, and parameters of lesser interest that simply need to be accounted for ("nuisance parameters").

In this construct, all observed quantities are fixed and are conditioned on, and all unobserved quantities are assumed to possess distributional qualities and therefore are treated as random variables. Thus, parameters are now no longer treated as fixed unmoving (like a classicist would assume, using the NHST) in the total population, and all statements are made in probabilistic terms.

Overview of Bayesian Inference

The Bayesian inference process starts with assigning *prior distributions* for the unknown parameters. These unknown parameters are operationalized with observed explanatory variables in a simple model. *Prior distributions* range from very informative descriptions based on previous research in the field to deliberately vague and uncertain forms that reflect high levels of uncertainty or previous ignorance. It is important to note that this prior distribution is not an inconvenience imposed by the treatment of unknown quantities. It is the means by which existing knowledge is systematically included in the model. Importantly, this prior information can include qualitative, narrative, statistical, and intuitive knowledge.

The second step in the process requires the specification of a likelihood function in the conventional manner by assigning a parametric form and plugging in the observed data. This is done in exactly the conventional likelihoodist fashion.[12] The researcher can choose the most appropriate parametric form, including a simple linear model as we do below.

The third step is to produce a *posterior distribution* by multiplying the prior distribution by the likelihood function. In this manner, the likelihood function uses the data to *update* the specified prior knowledge conditionally.

We can summarize probabilistic information about an unknown parameter according to:

$$\text{Posterior Probability} \propto \text{Prior Probability} \times \text{Likelihood Function.}$$

This is just Bayes' Law, wherein the denominator on the right-hand side has been ignored by using proportionality. The symbol "$\propto$" stands for "proportional to," which means that constants have been left out that make the posterior sum or integrate to one as is required of standardized probability mass functions and probability density functions. Renormalizing to a standard from can always be done later, plus, using proportionality is more intuitive and usually reduces the calculation burden. What this "formula," above, shows is that the posterior distribution is a compromise between the prior distribution, reflecting research beliefs, and the likelihood function, which is the contribution of the data at hand.[13]

The fourth step is to evaluate the fit of the model to the data and the sensitivity of the conclusions to the assumptions. This can be done systematically,[14] or in an *ad hoc* fashion by trying different reasonable priors or likelihood functions.[15]

When the researcher is happy with the fit of the model and the range of assumptions, the results are described to readers. Unlike the Null Hypothesis Significance Test method of deciding strength of conclusions based on the magnitude of p-values, evidence is presented in the Bayesian inference process by simply summarizing the posterior distribution, and therefore there is no artificial decision based on the assumption of a true null hypothesis. Posterior summary is usually done with quantiles and probability statements such as the probability that the parameter of interest is less than/greater than some interesting cnstant, or the probability that this parameter occupies some region.

Note also that if the posterior distributions is now treated as a new prior distribution, it too can be improved if new data are observed. In this way, the Bayesian paradigm provides a means of scientifically updating knowledge about the parameters of interest that is updated and accumulated over time.[16] One scholar's results can them be incorporated into subsequent analysis. We now describe these steps in greater detail:

Specifying the Likelihood Function

Suppose collected data are treated as a fixed quantity and we know the appropriate probability mass function or probability density function for describing the data-generation process. Standard likelihood and Bayesian methods are similar in that they both start with these two suppositions and then develop estimates of the unknown parameters in the parametric

model. Maximum likelihood estimation substitutes the unbounded notion of likelihood for the bounded definition of probability by starting with Bayes' Law:

$$p(\beta \mid \mathbf{X}) = \frac{p(\beta)}{p(\mathbf{X})} p(\mathbf{X} \mid \beta), \tag{2.1}$$

where β is the unknown parameter of interest and X is the collected data. The key is to treat $\frac{p(\beta)}{p(\mathbf{X})}$ as an unknown function of the data independent of $p(\mathbf{X}|\beta)$. This allows us to use: $L(\beta|\mathrm{X}) \propto p(\mathrm{X}|\beta)$. Since the data are fixed, then different values of the likelihood function are obtained merely by inserting different values of the unknown parameter, β, or (more realistically) the parameter vector, β.[17]

The likelihood function, $L(\beta|\mathrm{X})$, is similar to the desired but unavailable inverse probability, $p(\mathrm{X}|\beta)$, in that it facilitates testing alternate values of β to find a most probable value: $\hat{\beta}$. Thus, interest is generally in obtaining the *maximum likelihood* estimate of β: the value of the unconstrained and unknown parameter, β, which provides the maximum value of the likelihood function, $L(\beta|\mathrm{X})$, denoted $\hat{\beta}$.

In this way, $\hat{\beta}$ is the most likely to have generated the data given a specific parametric form relative to other possible values in the sample space of β. However, since 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. Note that the prior, $p(\beta)$, is essentially ignored here rather than overtly addressed. This is equivalent to assigning a uniform prior in a Bayesian context, an observation that has led some to consider classical inference to be a special case of Bayesianism: "everybody is a Bayesian; some know it."

Applying the Prior to Obtain the Posterior

The Bayesian approach addresses the inverse probability problem by making distributional assumptions about the unconditional distribution of the parameter, β, *prior* to observing the data, X: $p(\beta)$. The prior and likelihood are joined with Bayes' Law:

$$\pi(\beta \mid \mathbf{X}) = \frac{p(\beta) L(\beta \mid \mathbf{X})}{\int_{\Theta} p(\beta) L(\beta \mid \mathbf{X}) d\beta} \tag{2.2}$$

to produce the posterior distribution, where $\int_{\Theta} p(\beta) L(\beta \mid \mathbf{X}) d\beta = p(\mathbf{X})$. Here the $\pi()$ notation is used to distinguish the posterior distribution for β from the prior. The very natural and intuitive interpretation of the posterior

distribution is that it tells us all that we know about β given the observed data, **X**, and other information summarized in the prior distribution.

The term in the denominator of the right-hand-side of the equation is generally not important in making inferences and can be recovered later by integration. This term is typically called the *normalizing constant*, the *normalizing factor*, or the *prior predictive distribution*, although it is actually just the marginal distribution of the data, and ensures that $\pi(\beta \mid \mathbf{X})$ integrates to one.

A more compact and useful form of the equation is developed by dropping this denominator and using proportional notation since $p(\mathrm{X})$ does not depend on β and therefore provides no relative inferential information about more likely values of β:

$$\pi(\beta \mid \mathbf{X}) \propto \rho(\beta) L(\beta \mid \mathbf{X}),$$

meaning that the unnormalized *posterior* (sampling) distribution of the parameter of interest is proportional to the prior distribution times the likelihood function.

The maximum likelihood estimate is equal to the Bayesian posterior mode with the appropriate uniform prior, and they are asymptotically (as the data size gets very large) equal given *any* prior: both are normally distributed in the limit. In many cases, the choice of a prior is not especially important since as the sample size increases, the likelihood progressively dominates the prior. While the Bayesian assignment of a prior distribution for the unknown parameters can be seen as subjective (though *all* statistical models are actually subjective), there often are 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 (conjugacy), and the prior allows researchers to include additional information collected outside the current study. A more detailed discussion concerning the use of the prior is made below.

Summarizing Bayesian Results

Bayesian researchers are generally not concerned with just getting a specific point estimate of the parameter of interest, β, as a way of providing empirical evidence in probability distributions. Rather, the focus is on describing the shape and characteristics of the posterior distribution of β. Such descriptions are typically in the form of credible intervals, quantiles of the posterior, and particular probabilities of interest such as $p(\beta < 0)$.

Credible intervals are the Bayesian equivalent of confidence intervals, in which the region describes the area over the coefficient's support with the

$(1-\alpha)\%$ of the density symmetrically around the posterior mean. Like frequentist confidence intervals, a credible interval that does not contain zero implies that the coefficient estimate is deemed to be reliable, but instead of being $(1-\alpha)\%$ "confident" that the interval covers the true parameter value, a credible interval provides a $(1-\alpha)\%$ probability that the true parameter is in the interval.

Sometimes credible intervals are relatively wide, indicating that the coefficient varies considerably and is less likely to be reliable, and sometimes these regions are quite narrow, indicating greater certainty about the central location of the parameter.

The final step in Bayesian analysis is to evaluate the model fit to the data and determine the sensitivity of the posterior distribution to the various assumptions made. Here, it is usual to modify these assumptions somewhat and observe whether the conclusions remain both statistically reliable *and* stable under such modest changes. Such changes include altering the prior specification and using alternate forms of the likelihood function.

Obtaining the Posterior Through Simulation

Markov Chain Monte Carlo (MCMC) techniques solve a lingering problem in Bayesian analysis. Often Bayesian model specifications that were either interesting or realistic produced inference problems that were analytically intractable. The basic principle behind MCMC techniques is that if an iterative chain of consecutive values can be set up carefully and run long enough, then *empirical* estimations of quantities of interest can be obtained from chain values. So to estimate multidimensional probability structures (i.e., desired posteriors), we start a Markov Chain in the appropriate sample space and let it run until it settles into the correct distribution. Then, when it runs for some time confined to this particular distribution, we can collect statistics such as means, variances, and quantiles from the simulated values.

The most common method of producing Markov chains for MCMC work is the Gibbs sampler, which produces an empirical estimate of the posterior distribution of interest by iteratively sampling from full conditional distributions. The result is an estimate of the coefficients that avoids difficult or impossible analytical calculations.[18] The Gibbs method is popular because of this conditional specification, that is, all one has to do is elaborate each conditional dependency.

The Gibbs sampler works in detail as follows.[19] For convenience define ϕ as a k-dimensional vector of unknown parameters. Call $\phi_{[i]}$ the ϕ vector where the i^{th} parameter is jackknifed from the vector (temporarily omitted). The Gibbs sampler draws from the complete conditional distribution for the "left out" value: $\pi(\phi_i \mid \phi_{[i]})$, repeating for each value in the vector each time conditioning on the most recent draw of the other parameters. When each of

the parameters has been updated in this way, then the cycle recommences with the completely new vector ϕ.

This procedure will converge permanently to a limiting (stationary) distribution that is the target posterior, provided that the chain is *ergodic*. A chain is ergodic if it is aperiodic and positive recurrent. Aperiodic chains have no defined "looping", whereby they repeat the same series of values in a given period. A Markov chain is recurrent if it is defined on an irreducible state space such that every substate can be reached from every other substate. A Markov chain is *positive* recurrent if the mean time to transition back to the same state is finite. The ergodic theorem is foundational to MCMC work. It is essentially the strong law of large numbers in a Markov Chain sense: the mean of chain values converge almost surely to strongly consistent estimates of the parameters of the limiting distribution, despite mild dependence (on some state space $S \in \Re$ for a given transition kernel and initial distribution). These properties for the Gibbs sampler are well known and not further discussed here.[20] The original article on the statistical application of Gibbs sampling[21] is a far more demanding read and applies the algorithm to photo image restoration.

Although the ergodic theorem shows that after a sufficiently large number of chain iterations are performed, subsequent draws are from the target limiting posterior distribution: $\pi(\phi \mid X)$, reality is rarely this clear, and the primary focus of the current MCMC literature is on assessing convergence. Two primary philosophies compete for adherents among applied researchers. Gelman and Rubin[22] suggest using the EM algorithm (or some variant) to find the mode or modes of the posterior, then create an overdispersed estimate of the posterior as a starting point for multiple chains. Convergence is assessed by comparing within-chain variance against between-chain variance with the idea that at convergence, variability within each chain should be similar and will resemble the estimated target variance.

Conversely, Geyer[23] recommends implementing one long chain and using well-known time series statistics to assess convergence. In practice, most researchers are not as canonical as either specified approach and perform some combination of them. The approach taken here is to run multiple chains during a burn-in period, assess convergence, and then, upon success, let one chain run longer. The burn-in period is an interval in which the Markov Chain is allowed to run without concern for its trajectory. The idea is to let the chain run for a sufficiently long period of time as to "forget" its starting point. If the chain reaches an equilibrium condition, it is moving through the state space of the target distribution and empirical draws of its position represent samples from the desired limiting distribution. So assessing convergence is vital to determining the quality of the resulting inferences.

Though the Gelman and Rubin diagnostic[24] is widely accepted, there are alternative diagnostic techniques. Geweke's[25] convergence statistic relies on a comparison of some proportion of the early part of the chain after the burn-in

period with some non-overlapping proportion of the late part of the chain. Geweke proposes a difference of means test using an asymptotic approximation of the standard error for the difference. Since the test statistic is asymptotically standard normal, then for reasonably long chains, small values imply that the chain has converged, which is quite intuitive. Conversely, values that are atypical of a standard normal distribution provide evidence that the two selected portions of the chain differ reasonably (in the first moment), and one then concludes that the chain has not converged. The selected window proportions can change the value of the test statistic if the chain has not converged. Therefore, a further diagnostic procedure involves experimenting with these proportions. The final reported values (0.1 and 0.5) are Geweke's default recommendation, but similar results were observed using close alternatives.

EMPIRICAL EXAMPLE: MEIER'S MODEL OF EDUCATIONAL EFFECTS

We illustrate the Bayesian model through a partial replication of a Meier, Polinard, and Wrinkle[26] study of the bureaucratic effects on education outcomes in public schools. Meier, et al. are concerned with whether the education bureaucracy is the product or cause of poor student performance. The issue is one of contention in the literature because of the implications of these conclusions on the school choice debate.[27] Chubb and Moe[28] have long argued that the institutional structure of the schools, especially the overhead democratic control, resulted in the schools being ineffective. The institutional structure and the bureaucracy have created a process that leads to poor performance by the public schools.[29] This conclusion was challenged by Meier and Smith,[30] who contend that bureaucracy is an adaptation to poor performance and not the cause.

The authors are seeking to identify the variables that cause poor scholastic performance. Though the methodologies are *presumably* objective, the debate on school choice is complex and can clearly be far more normative.[31] For the American public, the issue of school performance is particularly important, and the concern is often associated with funding issues.[32] Further, the nature of Chubb and Moe's[33] assertion about democratic control can raise a more theoretical debate about the role of democratic structures and public administration in society.[34] Some have argued that the role and efficacy of the public schools is measured in part by their place in civic society.[35] These implications are significant, not simply so that one can take sides, but rather because they demonstrate that the methodology chosen as well as the specification of any model are not made in an abstract value-free universe, but within existing context.

In addressing this conflict, Meier, Polinard, and Wrinkle use a linear model based on a panel dataset of over 1000 school districts for a seven-year

period to test organizational theory and educational policy. The authors use these data to find whether there is a causal relationship between bureaucracy and poor performance by public schools.[36] As noted above, the central issue in this literature is one of causality through a "production function" that maps inputs to outputs in essentially an economic construct. Along with bureaucracy, student and school performance can be influenced by a number of variables, some of which are causally related, including class size, state funding, teacher salary, and experience. Meier et al. measure bureaucracy as the total number of full-time administrators per 100 students and lag the variable so as to create a more likely causal relationship. The authors concede this measure is incomplete, as it captures only a portion of the more complex definition of bureaucracy.[37] Nonetheless, they claim the measure has "substantial face validity."[38]

The control variables for educational performance included by the authors reflect student characteristics, measures of resources and district policies. Specifically, they include three measures of financial capital, which consist of the average teacher salary, per pupil expenditures for instruction, and the percentage of money each district receives from state funds. A measure of human capital was included based on teacher experience, and two policy indicators were used by measuring the average class size in the district and the percent of students in gifted classes.[39] Though these explanatory variables are placed in separate categories, some clearly are measuring concepts that are difficult to distinguish.

The linear model proposed by the authors is affected by both serial correlation and heteroscedasticity. Meier et al. address these concerns through a set of six dummy variables for each year as well as through the use of a weighted least squares.[40] The authors discuss both issues at length, but conclude that though the heteroscedasticity was significant, it was ultimately trivial in consequence. We account for this in a better way by adding a random effects term to the Bayesian model that allows for greater individual heterogeneity of the error term, but in the same model specification.[41]

Meier et al. then present the finding that their lagged measure of bureaucracy does not have a significant influence on the outcome variable of student performance measured by the pass rate. Additionally, in the model, measures of poverty as well as the measures related to gifted classes and class size are significant at conventional levels using the NHST ($T > 4$).[42]

The Education Production Function

The dominant framework for describing education policies and outcomes is the *education production function*: how the mixture of inputs combines with the established processes to determine learning outcomes. *Essentially all model specifications in this literature (Bayesian or otherwise) reduce to different*

interpretations of the form of the education production function. This has been described in other language as "the relationship among the different inputs and into outcomes of the educational process."[43] This is essentially an economic model, which depicts student achievement as a direct function of resource inputs and the way those inputs are applied.[44] In this framework, a school is considered essentially the way economists view the firm: acquiring and managing inputs to process clients with the goal of maximizing some defined output.

If we subscribe to the standard economic conditions for technical efficiency as prescribed in the traditional literature, then a realistic description of the relationship between inputs and outputs appears even less obtainable. The required conditions typically include:[45]

1. administrators control input allocation;
2. competition for "customers";
3. knowledge of input and output pricing and availability;
4. an identifiable production process; and
5. accurate feedback on success or failure.

Control of input allocation obviously varies according to the level of the administrator, but it certainly is not absolute at *any* level. Competition for customers is an interesting and currently debated question. Some scholars have argued that treating the student body as customers can lead to a decline in the rigor of the curricula and teaching methods.[46] In general, public school administrators do not compete, nor do they consider students to be customers in the traditional economic sense. One notable exception is the genesis of charter schools, which are small in number but increasingly popular. Knowledge about pricing is probably the most reasonable of these assumptions in the public education context. However, it is clear that this knowledge is certainly not absolute (nor would it be in most commercial settings), and output pricing is not really an applicable concept. The idea that the education production function is identifiable to educators and administrators is greatly debated, but no scholar is willing to assert that public school managers in the aggregate have a substantial claim on the form of the production function.

THE MEIER MODEL: SUMMARY AND REPLICATION

We address this debate initially by replicating the exact linear model of Meier, Polinard, and Wrinkle,[47] except using an explicitly Bayesian setup. That is, we specified normal (but very diffuse) priors on the parameters and calculated a posterior distribution according to the Bayesian principle. These specifications reflect the underlying Gauss-Markov assumptions of the standard linear model, but are re-expressed in a Bayesian context as described in previous

sections. Obviously we did not have to do this since there is nothing methodologically wrong with the original linear model estimated with least squares (other than a Null Hypothesis Significance Test interpretation, perhaps). However, our Bayesian replication of the classicist result is done to solidify the integrity of our proposed paradigm and to serve as a starting point for the more elaborate Bayesian specifications to come.

Our estimation of the Meier et al. model is done using Gibbs Sampling as implemented in the WinBUGS package. This also was not essential since the normal priors are *conjugate*, meaning that the resulting posterior form is also normal and that the posterior can be analytically calculated. However, we are trying to demonstrate the flexibility of this computing approach as well as the general Bayesian perspective.

Since the first stage involved replicating a classicist or NHST model, we assumed no prior knowledge and incorporated that ignorance into the model in the form of diffuse Gaussian normal priors with a large variance. Another expression of prior ignorance is obtained by centering these distributions at zero.[48] A large variance gives a diffuse prior, which usually leads to a posterior that favors information from the likelihood function, unless the sample size is small.[49]

The Bayesian model specifications are often given in "stacked" notation that summarizes the distribution assumptions (priors and likelihood):

$$
\begin{aligned}
&Y[i] \sim N(\lambda[i], \sigma^2),\\
&\lambda[i] = \beta_0 + \beta_1 x_1[i] + \ldots + \beta_k x_k[i] + \varepsilon[i]\\
&\varepsilon[i] \sim N(0.0, \tau)\\
&\beta[i] \sim N(0.0, 10)
\end{aligned}
$$

for $i = 1{:}\ n$. This means that the outcome variable is assumed to be normally distributed around the systematic component with fixed variance (line 1), this systematic effect (λ) is a linear additive specification with a random effects term (lines 2 and 3), and that the coefficient estimates are given identical diffuse normal priors (line 4). The explanatory variables are denoted x_i and are columns from the explanatory variable matrix X.

Additionally, we corrected for missing data, not by deletion, by case, but by imputing missing data points through Bayesian estimation using the R package: *multiple imputation by chained equations* (mice). Multiple imputation creates a posterior distribution for the missing data conditional on the observed data, then draws randomly from this distribution to create multiple replications (5–10) of the original dataset, each of which are then analyzed. Final summary of the results is produced by an average of the resulting coefficients (with a standard error adjustment).[50]

Table 1. Posterior Summary, Meier Replication Model

Explanatory Variables	Mean Effects	Standard Deviation	95% Credible Interval
Constant Term	9.172	1.358	[6.510: 11.840]
Lag of Student Pass Rate	0.677	0.008	[0.661: 0.693]
Lag of Bureaucrats	−0.081	0.262	[−0.595: 0.431]
Low Income Students	−0.108	0.006	[−0.119: −0.097]
Teacher Salaries	0.073	0.053	[−0.035: 0.181]
Teacher Experience	−0.009	0.046	[−0.099: 0.082]
Gifted Classes	0.097	0.023	[0.054: 0.139]
Class Size	−0.220	0.052	[−0.322: −0.118]
State Aid Percentage	−0.002	0.004	[−0.010: 0.006]
Funding Per Student (×1000)	0.065	0.174	[−0.276: 0.406]
	Posterior standard error of $\tau = 0.00072$		

Meier, Polinard, and Wrinkle argue that their results are robust, and that using Bayesian estimation with diffuse priors does not change this assessment or the statistical outcome. We were able to replicate the original study with very little difference (the small changes were likely due to our imputation of the missing data and a different treatment by the original authors). A table of our output is included in Table 1. The replicated linear model here is (unsurprisingly) very robust, and it supports the original conclusions of the authors. Specifically, we also find support for the demographic variables as well as class size effects.

INSERTING SUBSTANTIVE PRIOR INFORMATION

The Bayesian Model with Meier-Priors

As illustrated above, the Meier et al. model is robust to modest changes in the assumptions. With a model that is particularly stable, there initially appears to be little that a Bayesian regression analysis can add. Nonetheless, there has been extensive work on related subjects concerning the role of bureaucracy as well as other factors in the performance of public schools. In fact, scholars have done extensive work in the area.[51] Bayesian estimation allows us to incorporate this previous work within the new model.

To expand on the Meier et al. model, we included non-sample information for the creation of the Bayesian prior drawn from Meier's previous work on school bureaucracy and school performance with Kevin Smith.[52] Clearly, Meier et al. were not uninformed entering their more recent study,[53] and

Bayesian inference allows for the incorporation of that knowledge. The Smith and Meier (1995) work includes data and inference on the impact of funding and other institutional variables on student achievement in Florida. These data include district level data for all of the public schools in Florida. Smith and Meier note that the Florida data provides a diverse group of students with constant measures over time. The Florida data represents both rural and urban districts as well as different ethnic and socioeconomic compositions.[54]

The prior information was added to the 2000 Meier, Polinard, and Wrinkle study by incorporating the previous study's findings into the distributions for the explanatory variables, creating "Meier-priors." The distributions remain normal, but are now centered around values drawn from 1995 Smith and Meier findings. Thus, the model specification differs from the replication only in the prior assumptions for the β[55]:

$\beta[0] \sim N(0.0,10)$ $\beta[1] \sim N(-0.025,10)$ $\beta[2] \sim N(0.0,10)$ $\beta[3] \sim N(0.23,10)$

$\beta[4] \sim N(0.615,10)$ $\beta[5] \sim N(-0.068,10)$ $\beta[6] \sim N(0.0,10)$ $\beta[7] \sim N(-0.033,10)$

$\beta[8] \sim N(0.299,10)$ $\beta[9] \sim N(0.0,10)$ $\beta[10] \sim N(0.0,10)$ $\beta[11] \sim N(0.0,10)$

$\beta[12] \sim N(0.0,10)$ $\beta[13] \sim N(0.0,10)$ $\beta[14] \sim N(0.0,10)$ $\beta[15] \sim N(0.0,10)$

where obviously some of these are left uninformed. The data drawn from the Smith and Meier research were insufficient to address all of the variables in the current models. For variables without adequate information to form a more specific prior, a diffuse prior centered at zero was used. This permits the prior to be largely shaped by the data.[56] Additional research in the field may allow for a greater degree of specification for the unknown variables.

Meier et al. are relying on the linear model assumptions of asymptotic normality and constant variance of the residuals, whether they state it or not. Our model employs a random effects specification with the term τ specified as an inverse-gamma distribution for the variance in the outcome variable allowing for greater unit heterogeneity than the Meier model, but in the same specification for a random effects model. The inverse-gamma distribution is appropriate because it is the conjugate prior for the normal-linear model variance. Instead of having a constant variance, the model will draw on the inverse-gamma distribution for that parameter. This makes the MCMC estimation procedure work much better, and it allows for systematic inclusion of data heterogeneity.

In the Meier-prior model, as in the case of the initial simple replicated model, the lagged burcaucracy variable was not significant in the new Bayesian result. However, even in this case, where the model was particularly robust, the use of the Bayesian estimation illustrates that the initial NHST model was incomplete. The impact of teacher salaries on student performance is significant in these data despite the original finding in the Meier et al. linear model. Our results are displayed in Table 2.

Table 2. Posterior Summary, Meier-Prior Model

Explanatory Variables	Mean Effects	Standard Deviation	95% Credible Interval
Constant Term	9.173	1.358	[6.510: 1JL1.840]
Lag of Student Pass Rate	0.686	0.008	[0.670: 0.701]
Lag of Bureaucrats	0.004	0.258	[–0.498: 0.516]
Low Income Students	–0.149	0.006	[–0.116: –0.094]
Teacher Salaries	0.196	0.053	[0.920: 0.300]
Teacher Experience	–0.549	0.046	[–0.144: 0.035]
Gifted Classes	0.096	0.021	[0.055: 0.138]
Class Size	–0.216	0.051	[–0.316: –0.116]
State Aid Percentage	0.004	0.004	[–0.004: 0.012]
Funding Per Student (×1000)	0.118	0.172	[–0.219: 0.455]
Posterior standard error of $\tau = 0.00072$			

Interestingly, Meier et al. expected to find a positive relationship between teacher salaries and student performance.[57] The original NHST model was unable to find the positive link between the salaries and the test scores. This is true despite both the expectations of the authors as well as the prior research that suggested such a relationship should exist.[58] The posterior distribution of the model indicates a positive relationship with a 95% credible interval bounded away from zero: [–0.116:–0.094].

Our result does not actually reject the findings of Meier et al., nor are we suggesting that the conclusions reached by those scholars are in error. In fact, the Bayesian model generated results that were closer to the expectation of the researchers since it incorporated knowledge to which the researchers already had access. Meier et al. had noted that economic theory provides that higher salaries attract better teachers.[59] Hence, the incorporation of prior research, especially a scholar's own work, should be the norm, especially in the area of public administration where more practical and germane prescriptive findings are sought.

We also replicated the model using an interaction between class size and teacher salary as the variables are related.[60] The information for the Bayesian priors were again drawn from the previous research by Smith and Meier.[61] The interaction is multiplicatively defined, and given its own prior. As in the case of some of the other explanatory variables, since we had no prior expectations for the sign or magnitude of such an interaction, the interaction coefficient is assigned a normal prior centered at zero with a large variance.

Interestingly, the interaction coefficient was found to have a negative sign, with 95% credible interval bounded away from zero, as provided in Table 3. This means that larger class sizes have a dampening effect on the

Table 3. Posterior Summary, Interaction Model

Explanatory Variables	Mean Effects	Standard Deviation	95% Credible Interval
Constant Term	4.799	2.373	[0.165: 9.516]
Lag of Student Pass Rate	0.684	0.008	[0.667: 0.699]
Lag of Bureaucrats	–0.042	0.261	[–0.557: 0.469]
Low Income Students	–0.105	0.006	[–0.117: –0.094]
Teacher Salaries	0.382	0.099	[0.189: 0.575]
Teacher Experience	–0.066	0.046	[–0.156: 0.025]
Gifted Classes	0.096	0.021	[0.054: 0.138]
Class Size	0.196	0.191	[–0.180: 0.569]
State Aid Percentage	0.002	0.004	[–0.006: 0.010]
Funding Per Student (×1000)	0.049	0.175	[–0.294: 0.392]
Class Size × Teacher Salaries	–0.015	0.007	[–0.029: –0.002]
Posterior standard error of $\tau = 0.00071$			

positive (reliable) impact of increasing teacher salaries. Interestingly, the posterior distribution for class size now shows a much less reliable effect in the interaction model (the 95% credible interval is nearly centered at zero). Interaction effects can sometimes be "kleptomaniacs" in that they serially steal explanatory power and reliability from main effects. This finding says that the effects of class size are now only reliable in this model in the context of specified teacher salary levels. Notice that this changes the context of the education production function conceptualization considerably from the original specification.

The Bayesian Model with Chubb and Moe Priors

To observe the sensitivity of the model to competing research, we also created a model with prior information drawn from the work of Chubb and Moe.[62] As noted earlier, Chubb and Moe contend that political institutions place a burdensome bureaucracy on the public schools, which ultimately harms student achievement. The authors support their contention with statistical inferences based on an extensive survey of American high school students. Our new contrasting Bayesian model is still linear with normal distributions, but the prior values are drawn from the findings from Chubb and Moe instead of being taken from Smith and Meier, or left diffuse to operationalize ignorance. The resulting model specification differs from the previous models only in the prior assumptions for the β. As in the previous models, where the variables are

without sufficient information to form a more specific prior, a diffuse prior centered at zero was used.

$$\beta[0] \sim N(0.0,10) \quad \beta[1] \sim N(-.023,10)\ \beta[2] \sim N(0.025,10)\ \beta[3] \sim N(0.0,10)$$
$$\beta[4] \sim N(0.042,10)\ \beta[5] \sim N(0.016,10)\ \beta[6] \sim N(0.002,10)\ \beta[7] \sim N(-0.007,10)$$
$$\beta[8] \sim N(0.032,10)\ \beta[9] \sim N(-0.017,10)\ \beta[10] \sim N(0.0,10)\ \beta[11] \sim N(0.0,10)$$
$$\beta[12] \sim N(0.0,10) \quad \beta[13] \sim N(0.0,10) \quad \beta[14] \sim N(0.0,10) \quad \beta[15] \sim N(0.0,10)$$

The results from the Chubb and Moe model actually illustrate the particular strength of the Meier study. Despite unfavorable prior information, the resulting posteriors are not substantively affected as illustrated in Table 4, where the posterior distributions are very similar to those produced by the Meier-priors. Since the posterior distribution is a compromise between the prior distributions (our operationalizing of Chubb and Moe's 1990 research beliefs), and the likelihood function (contribution of the data at hand), then these findings indicate that the data are strongly informed about the structure of the specified education production function. Therefore, the last result confirms the general reliability of the findings by Meier et al. as the data is clearly dominating the resulting posterior, and therefore suggest that the controversy between these authors in the literature is somewhat misguided as to its focus.

Markov Chain Convergence

As noted above, the reliability of a posterior generated through MCMC is based upon an assumption of convergence. It is essential for the Markov chain

Table 4. Posterior Summary, Chubb-Moe Priors Model

Explanatory Variables	Mean Effects	Standard Deviation	95% Credible Interval
Constant Term	9.173	1.358	[6.510: 11.840]
Lag of Student Pass Rate	0.686	0.008	[0.670: 0.702]
Lag of Bureaucrats	0.004	0.259	[−0.506: 0.512]
Low Income Students	−0.105	0.006	[−0.116: −0.094]
Teacher Salaries	0.197	0.053	[0.091: 0.301]
Teacher Experience	−0.055	0.046	[−0.143: 0.034]
Gifted Classes	0.096	0.021	[0.054: 0.139]
Class Size	−0.216	0.051	[−0.314: −0.116]
State Aid Percentage	0.004	0.004	[−0.004: 0.001]
Funding Per Student (×1000)	0.118	0.174	[−0.223: 0.459]
Posterior standard error of $\tau = 0.00072$			

to have reached its stationary distribution for the resulting empirical summaries to be meaningful. We found that the specified model rapidly converged and mixed well throughout this distribution after approximately 20,000 iterations. Nonetheless, we ran the models much longer to be confident of convergence before reporting posteriors. As we discussed at length above, there are several different tests for convergence. Initially, we used the widely accepted Gelman and Rubin test.[63] Gelman and Rubin's convergence diagnostic uses a measure of within chain variance and between chain variance and is accessible with little difficulty through the WinBUGS (see Appendix) package. A score of 1.2 or less is considered acceptable for lacking evidence of non-convergence, and all of our subchains achieved a score of approximately.

For additional evidence to support convergence, we also used the Geweke test,[64] also discussed earlier. The Geweke test is essentially a t-test of means-difference with large enough sample size that the statistic is normally distributed with tail values supporting a difference in chain periods and therefore non-convergence. It is common in this context to rely upon the artificial, but convenient cut-off p-value of 0.05 when making such decisions. We implemented the Geweke test in R using the package: Bayesian Output Analysis (BOA, freely available at: www.public-health.uiowa.edu/boal/). BOA produces a menu driven interface for use with the statistical programs S-Plus and R. The p-values produced in our Geweke test were greater than 0.05 for each parameter sub-chain. We therefore find that the results do not provide any evidence against convergence.

DOES PRIOR MEAN BETTER?

The Bayesian approach has been criticized for the interjection of prior information that is subjective in nature. This subjectivity can be introduced, not only from the nature of the previous data used, but also by the weight given it in the specification of the prior distribution.[65] Arguably, the use of prior information will hurt the objective results obtained through the sample data alone. Criticism of the use of prior information is not new, and scholars have claimed that the use of this information would make the resulting experiment dependent on the interpretation of prior experience.[66] Similarly, claims have been made that the method itself does not reassure scholars that the data has been interpreted fairly.[67] Others have claimed that the use of priors will tempt scholars into selecting information based on a desired outcome rather than science: that Bayesian methods would provide a "larger means for skull-duggery."[68]

The nature of the criticism is based on assumptions about the objectivity of the NHST method that are simply untrue. In the NHST model, the prior is unspecified, but "flat" by default, representing a type of ignorance even when the researcher is bringing considerable knowledge to the study. We illustrated

this in our first replication by generating Meier model results in a Bayesian regression analysis simple by using a diffuse normal prior centered at zero. (See Table 1). No matter which methodology is employed, the researcher is making an assumption about prior ignorance each time she runs a model. This is true whether that assumption is accurate or not.

Secondly, a different but important type of prior information is regularly used in the standard NHST model through choices made about coding schemes, transformations, and interpretation. In the 2000 Meier et al. study, the authors make an assumption about the coding of bureaucracy based upon assumptions and experience they bring to the research. They measure bureaucracy as the total number of full-time administrators per 100 students and lag the variable so as to create a more likely causal relationship. Though this is admittedly incomplete, the authors claim the measure still has considerable information.[69] Others may not agree that a measure of administrators captures the key components of bureaucracy: that the authors' measure fails to account for rules, procedural restraints, or even the more ubiquitous notion of red tape, which is itself the subject of significant scholarly work.[70] Indeed, Meier et al. concede the omission,[71] but proceed nonetheless.

In the context of the school debate, the process of model specification is particularly influenced by a more normative context. It is one choice to structure an inquiry with a judgment about funding issues.[72] Chubb and Moe's contentions concerning democratic control raise a complex set of questions about the role of democratic structures and public administration in society beyond more basic issues of funding or average number of administrators.[73] Additionally, centering schools within a model of civic society presents a different set of variables to consider in specifying a model.[74] Making the issue one of school choice presents additional considerations.[75] No matter how objective one attempts to make a model, the methodology chosen as well as the specification of any model are not made in an abstract value free universe, but within existing substantive context.

Further, Meier et al. make additional assumptions about the relationship of the variables to each other as well as the expected outcome based on previous knowledge. Though each of the explanatory variables are measured separately in the Meier, et al. linear model, the authors concede that they are intended to capture three general concepts: student characteristics, measures of resources, and district policies.[76] Though not addressed at length by the authors, conceptually the nature of these related variables creates difficulty for the NHST model because closely related variables may carry little independent knowledge.[77]

The use of previous research through the Bayesian prior can address this problem and allow for a greater distinction in collinear or near-collinear variables. A Bayesian regression addresses issues of collinearity by using more individual coefficient information than an ordinary least squares model, which results in smaller standard errors for the regression coefficients.[78] As a result, the data is strengthened by the Bayesian model and this often generates more

precise coefficients. In formulating, specifying, and finalizing models, previous knowledge is not only used, it is vital. This is as true in a NHST model as in a Bayesian one; the key difference is that scholars using Bayesian approaches *admit it*.

Meier et al. openly discuss some of the assumptions that are inherent in their coding schemes, but many authors do not. Bayesian estimation forces scholars to display the assumptions concerning prior information within the context of the model and to justify the usage of that prior information.[79] Researchers employing the NHST model use prior information and may conceal it under subjective schema. Bayesians candidly bring the information to the model and allow the reader to determine whether its use and weight were sufficiently justified within the context of the study.

The incorporation of prior information may be necessary where there are simply too few data points for the asymptotic assumptions for a NHST model to work. Bayesian methods have been advocated for research involving a small number of observations and cases with nonstochastic data as it allows for estimates and predictions when there is insufficient data to fit the desired model using standard methods (Western and Jackman[80] produce an informative model with only 20 cases!). Small data sets that produce fragile statistical inference in a standard model are more effectively handled by the Bayesian approach because of the incorporation of prior information in the estimation. In non-Bayesian estimations, small data sets and collinearity will produce results that are highly sensitive to model specifications. This may result in researchers discarding a study, not because the data fails to provide information, but, rather, because the methodological tools are insufficient.

We concede that the use of prior information is subjective and calls for the researcher to make specifications that are subject to criticism, but note that the Bayesian is always responsible for overtly declaring and defending her assumptions. The weight given the priors in our replication was not great, and as a result, the findings are largely similar to the original study. However, our methods and assumptions are transparent and other scholars can make different assumptions about the prior research and justify those assumptions in a new model with smaller variances and less diffuse priors. The end result is a dialogue, rather than a defensiveness argument concerning the hidden assumptions and specifications of a NHST model.

CONCLUSION: A NEW DIRECTION?

The purpose of this article was to propose the Bayesian inferential process for statistical analysis in public administration research. We have demonstrated several of the advantages, including: overt statement of assumptions, the ability to include prior information, the reliance upon information in the data above other model components, and the flexibility of specifications.

We have also illustrated that the use of the default flat or diffuse prior in an NHST model to indicate ignorance is an assumption that can inadvertently alter results even in the most stable or robust of linear models, whether the researcher is aware of it or not. Bayesian model uncertainty is certainly characterized subjectively within the context of data collection and analysis, but so is every other approach. Conversely, the openness and flexibility of the Bayesian approach allows the researcher to be more sensitive to the types of data collected, including data that might be the result of convenience samples rather than the random sampling required for standard statistical methods. The Bayesian paradigm provides a means of scientifically updating knowledge about the parameters of interest.

The use of the standard quasi-frequentist NHST approach is the dominant methodology in the discipline.[81] Nonetheless, public administration research often involves the extensive use of prior information. There is little sense in continuing to rely upon a methodology that cannot account for previous research, even when that research has been performed by the same scholar. This is particularly true in the field of public administration, which has a rich descriptive history. Such prior knowledge and research can, and should be updated with empirical work rather than isolated as stand-alone findings.

REFERENCES

1. Gill, J; Meier, K. Public Administration Research and Practice: A Methodological Manifesto. Journal of Public Administration Research and Theory **2000**, 10 (1), 157–200.
2. Eulau, H. The Interventionist Thesis in The Place of Policy Analysis in Political Science: Five Perspectives. American Journal of Political Science **1977**, 31, 419–423.
3. Gill, J. Whose Variance is it Anyway? Interpreting Empirical Models with State-Level Data. State Politics and Policy Quarterly **2001**, 1, 313–338.
4. Freedman, D.A.; Lane, D. Significance Testing in a Nonstochastic Setting. In *A Festschrift for Eric L. Lehman*; Bickel, P. J.; Doksum, K. A.; and Hodges, J.L., Eds.; Wadsworth: Belmont, CA, 1993.
5. Lehmann, E. L.; Casella, G. *Theory of Point Estimation*; Springer-Verlag: New York, 1998.
6. Western, B.; Jackman, S. Bayesian Inference for Comparative Research. The American Political Science Review **1994**, 88 (3), 412–423.
7. Leamer, E. E. Lets Take the Con Out of Econometrics. American Economic Review **1983**, 73 (1), 31–43.
8. Hanushek, F. A. The Economics of Schooling: Production and Efficiency in Public Schools. Journal of Economic Literature **1986**, 24, 1141–1177.
9. Leamer, E. E. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*; John Wiley & Sons: New York, 1978.

10. Barnett, V. *Comparative Statistical Inference*; John Wiley & Sons: New York, 1973.
11. Box, G.E.P.; Tiao, G.C. *Bayesian Inference in Statistical Analysis*; Wiley & Sons: New York, 1973.
12. King, G. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*; Cambridge University Press: Cambridge, 1989.
13. Robert, C.P. *The Bayesian Choice: A Decision Theoretic Motivation*, 2nd Ed.; Springer-Verlag: New York, 2001.
14. Gill, J. *Bayesian Methods: A Social and Behavioral Sciences Approach*; Chapman & Hall: New York, 2002.
15. Leamer, E. E. *Specification Searches:* Ad Hoc *Inference with Nonexperimental Data*; John Wiley & Sons: New York, 1978.
16. Bernardo, J.M. and Smith, A.F.M. *Bayesian Theory*; John Wiley & Sons: New York, 1994.
17. King, G. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*; Cambridge University Press: Cambridge, 1989.
18. Jackman, S. Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo. American Journal of Political Science **2000**, 44 (2), 375–404.
19. Geman, S.; Geman, D. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence **1984**, 6, 721–741; Smith, A.F.M.; Roberts, G.O. Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods (with discussion). Journal of the Royal Statistical Society, Series B **1993**, 55, 3–24.
20. For a comprehensive discussions of the theory and practice of Gibbs sampling and MCMC in general see: Carlin, B.P.; Louis, T.A. *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd Ed.; Chapman & Hall: New York, 2000; Gamerman, D. *Markov Chain Monte Carlo*; Chapman & Hall; New York, 1997; Gelfand, A.E.; Smith, A.F.M. Sampling-based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association **1990**, 85, 389–409; Gelman, A.; Rubin, D.B. Inference from iterative simulation Using Multiple Sequences. Statistical Science **1992**, 7, 457–511; Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Chapman & Hall: New York, 1995; Geweke, J. Bayesian Inference in Econometric Models Using Monte Carlo Integration. Econometrica **1989** 57, 1317–1339; Gill, J. *Bayesian Methods: A Social and Behavioral Sciences Approach*; Chapman & Hall: New York, 2002; Tanner, M.A. *Tools for Statistic Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*; Springer-Verlag: New York, 1996.
21. Geman, S.; Geman, D. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence **1984**, 6, 721–741.

22. Gelman, A.; Rubin, D.B. Inference from interative simulation Using Multiple Sequences. Statistical Science **1992**, 7, 457–511.
.23. Geyer, C.J. Practical Markov Chain Monte Carlo. Statistical Science **1992**, 7, 473–511.
24. Gelman, A.; Rubin, D.B. Inference from iterative simulation Using Multiple Sequences. Statistical Science **1992**, 7, 457–511.
25. Geweke, J. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In *Bayesian Statistics 4*; Bernardo, J.M., Smith, A.F.M., Dawid, A.P., and Berger, J.O., Eds.; Oxford University Press: Oxford, 1992; 169–193.
26. Meier, K.J.; Polinard, J.L.; Wrinkle, R. Bureaucracy and Organizational Performance: Causality Arguments about Public Schools. American Journal of Political Science **2000**, *44* (3), 590–602.
27. Smith, K.B.; Meier, K.J. *The Case Against School Choice: Politics, Markets, and Fools*; M.E. Sharpe: Armonk, NY, 1995.
28. Chubb, J.E.; Moe, T. Politics, Markets and the Organization of American Schools. American Political Science Review **1988**, *82* (4), 1065–1089.
29. Chubb, John E.; Moe, Terry. *Politics, Markets and America's Schools*; Brookings Institution: Washington, D.C., 1990.
30. Meier, K.J.; Smith, K.B. Representative Democracy and Representative Bureaucracy. Social Science Quarterly **1994**, 75 (4), 798–803.
31. Smith, K.B.; Meier, K.J. *The Case Against School Choice: Politics, Markets, and Fools*; M.E. Sharpe: Armonk, NY, 1995.
32. Paul, P. Education Reform. American Demographics **2002**, *24* (8), 20–22.
33. Chubb, J.E.; Moe, T. *Politics, Markets and America's Schools*; Brookings Institution: Washington, D.C., 1990.
34. Ventriss, Curtis. Radical democratic thought and contemporary American public administration. American Review of Public Administration **1998**, *28* (3), 227–245.
35. Denhardt, J.V.; Hamilton, L.K. Community v. Self-interest: Citizen Perceptions of Schools as Civic Investments. Journal of Public Administration Research and Theory **2002**, *12* (1), 103–128.
36. Meier, K.J.; Polinard, J.L.; Wrinkle, R. Bureaucracy and Organizational Performance: Causality Arguments about Public Schools. American Journal of Political Science **2000**, *44* (3), 590–602.
37. Downs, A. *Inside Bureaucracy*; Little and Brown: Boston, 1967.
38. Meier, K.J.; Polinard, J.L.; Wrinkle, R. Bureaucracy and Organizational Performance: Causality Arguments about Public Schools. American Journal of Political Science **2000**, *44* (3), 590–602; 593.
39. Ibid.
40. Ibid.
41. Beck, N.; Katz, J.N. Random Coefficient Models for Time-Series-Cross-Section Data: The 2001 Version. Available at the Political Methodology Working Paper Archive: http://web.polmeth.ufl.edu/,2001.

42. Meier, K.J.; Polinard, J.L.; Wrinkle, R. formance: Opcit, p. 595.
43. Hanushek, E.A. The Economics of Schooling: Production and Efficiency in Public Schools. Journal of Economic Literature **1986**, 24, 1141–1177; 1148.
44. Hanushek, Eric A. The Economics of Schooling: Production and Efficiency in Public Schools. Journal of Economic Literature **1986**, 24, 1141–1177; Monk, D. Education Productivity Research: An Update and assessment of its Role in Education Finance Reform. Educational Evaluation and Policy Analysis **1992**, *14* (4), 307–332; Ferguson, R.F. Paying for Public Education: New Evidence on How and Why Money Matters. Haryard Journal on Legislation **1991**, 28, 465–498; Ferguson, R.F.; Ladd, H.F. How and Why Money Matters: An Analysis of Alabama Schools. In *Holding Schools Accountable: Performance-based Reform in Education*; Ladd, H. F., Ed.; Brookings Institution Press: Washington, D.C., 1996.
45. Aigner, D.J.; Chu, S.F. On Estimating the Industry Production Function. American Economic Review **1968**, 58, 826–837; Comanor, W.S.; Leibenstein, H. Allocative Efficiency, X-Efficiency, and the Measure of Welfare Losses. Economica **1969**, 26, 304–309; Timmer, C.P. Using Probabilistic Frontier Production Function to Measure Technical Efficiency. Journal of Political Economy **1971**, *79* (4), 776–794.
46. Carlson, P.M.; Fleisher, M.S. Shifting Realities in Higher Education: Today's Business Model Thereatens Our Academic Excellence. International Journal of Public Administration **2002**, 25, 1097–1111.
47. Meier, Kenneth J.; Polinard, J.L.; Wrinkle, R. Bureaucracy and Organizational Performance: Causality Arguments about Public Schools. American Journal of Political Science **2000**, *44* (3), 590–602.
48. Western, B.; Jackman, S. Bayesian Inference for Comparative Research. The American Political Science Review **1994**, *88* (3), 412–423.
49. Lee, P.M. *Bayesian Statistics: An Introduction*; Oxford University Press: New York, 1989; Pollard, W.E. *Bayesian Statistics for Evaluation Research: An Introduction*; Sage: Beverly Hills, 1986.
50. For theoretical details review: Rubin, D. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons: New York, 1987; Little, Roderick J.A. and Rubin, D.B. On Jointly Estimating Parameters and Missing Data by Maximizing the Complete-Data Likelihood. The American Statistician **1983**, 37, 218–220.
51. Hanushek, E.A. The Economics of Schooling: Production and Efficiency in Public Schools. Journal of Economic Literature **1986**. 24, 1141–1177.
52. Smith, K.B.; Meier, K.J. *The Case Against School Choice: Politics, Markets, and Fools*; M.E. Sharpe: Armonk, NY, 1995.
53. Meier, K.J.; Polinard, J.L.; Wrinkle, R. Bureaucracy and Organizational Performance: Causality Arguments about Public Schools. American Journal of Political Science **2000**, *44* (3), 590–602.
54. Smith, K.B.; Meier, K.J. *The Case Against School Choice: Politics, Markets, and Fools*; M.E. Sharpe: Armonk, NY, 1995; 44.

55. The term β was used for ease of computation and reference. Each numbered β represents the prior information assigned to each explanatory variable. For reference: beta[0]=constant; beta[1]=Low Income Students; beta[2]=Teacher Salaries; beta[3]=Teacher Experience; beta[4]=Gifted Classes; beta[5]=Class Size; beta[6]=Percent State Funding; beta[7]= Funding Per Student; beta[8]=Lag of Student Pass Rate; beta[9]=Lag of Beaucrats; betas[10-15] represent the years 1993–1997.
56. Gill, J. *Bayesian Methods: A Social and Behavioral Sciences Approach*; Chapman & Hall: New York, 2002.
57. Meier, K.J.; Polinard, J.L.; Wrinkle, R. Bureaucracy and Organizational Performance: Causality Arguments about Public Schools. American Journal of Political Science **2000**, *44* (3), 590–602; 594.
58. Hanushek, E.A.; Pace, R. Who Chooses to Teach (and Why)? Economics of Education Review **1995**, *14* (1), 107–117.
59. Meier, K.J.; Polinard, J.L.; Wrinkle, R. Opcit, p. 593.
60. The prior for the interaction variable was operationalized in the model as beta [16].
61. Smith, K.B.; Meier, K.J. *The Case Against School Choice: Politics, Markets, and Fools*; M.E. Sharpe: Armonk, NY, 1995.
62. Chubb, J.E.; Moe, T. *Politics, Markets and America's Schools*; Brookings Institution: Washington, D.C., 1990.
63. Gelman, A.; Rubin, D. B. Inference from iterative simulation Using Multiple Sequences. Statistical Science **1992**, 7, 457–511.
64. Geweke, J. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In *Bayesian Statistics 4*; Bernardo, J.M.; Smith, A.F.M.; Dawid, A.P.; Berger, J.O., Eds.; Oxford University Press: Oxford, 1992; 169–193.
65. Carlin, B.P.; Louis, T.A. *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd Ed.; Chapman & Hall: New York, 2000.
66. Fisher, R.A. *The Design of Experiments*; Oliver and Boyd: London, 1935.
67. Efron, B. Why Isn't Everyone a Bayesian? American Statistician **1986**, 40, 1–11.
68. Firebaugh, G. Will Bayesian Inference Help? A Skeptical View. Sociological Methodology **1991**, 25, 469–492; 471 pp.
69. Meier, K.J.; Polinard, J.L.; Wrinkle, R. Bureaucracy and Organizational Performance: Causality Arguments about Public Schools. American Journal of Political Science **2000**, *44* (3), 590–602; 593.
70. Pandey, S.K. and Scott, P.G. Red tape: A Review and Assessment of Concepts and Measures. Journal of Public Administration Research and Theory **2002**, *12* (4), 553–581.
71. Meier, K.J.; Polinard, J.L.; Wrinkle, R. Opcit, 590–602. See also: Downs, A. *Inside Bureaucracy*; Little and Brown: Boston, 1967.
72. Paul, P. Education Reform. American Demographics 2002, *24* (8), 20–22.

73. Ventriss, C. Radical democratic thought and contemporary American Public administration. American Review of Public Administration **1998**, *28* (3), 227–245.
74. Denhardt, J.V.; Hamilton, L.K. Community v. self-interest: Citizen Perceptions of Schools as Civic Investments. Journal of Public Administration Research and Theory **2002**, *12* (1), 103–128.
75. Smith, K.B.; Meier, K.J. *The Case Against School Choice: Politics, Markets, and Fools*; M.E. Sharpe: Armonk, NY, 1995.
76. Meier, K.J.; Polinard, J.L.; Wrinkle, R. Bureaucracy and Organizational Performance: Causality Arguments about Public Schools. American Journal of Political Science **2000**, *44* (3), 590–602.
77. Western, B.; Jackman, S. Bayesian Inference for Comparative Research. The American Political Science Review **1994**, *88* (3), 412–423.
78. Box, G.E.P.; Tiao, G.C. *Bayesian Inference in Statistical Analysis*; Wiley & Sons: New York, 1973.
79. Gill, J. *Bayesian Methods: A Social and Behavioral Sciences Approach*; Chapman & Hall: New York, 2002; Ch. 5.
80. Western, B.; Jackman, S. Opcit.
81. Scott, P.G.; Falcone, S. Comparing public and private organizations. American Review of Public Administration **1998**, *28* (2), 126–145.

APPENDIX: WINBUGS CODE FOR THE SPECIFIED MODELS

Replication Model. This WinBUGS code replicates the Meier, Polinard, and Wrinkle (2000) finding using a Bayesian approach.

```
Model {
    beta0~dnorm(0.0,0.001);   beta1~dnorm(-20,0.001)I(-10,0);   beta2~dnorm(0.0,0.001
    beta3~dnorm(0.0,0.001);   beta4~dnorm(0.0,0.001);           beta5~dnorm(-0.0,0.00
    beta6~dnorm(0.0,0.001);   beta7~dnorm(-0.0,0.001)           beta8~dnorm(0.0,0.001
    beta9~dnorm(0.0,0.001);   beta10~dnorm(0.0,0.001);          beta11~dnorm(0.0,0.00
    beta12~dnorm(0.0,0.001);  beta13~dnorm(0.0,0.001)           beta14~dnorm(0.0,0.00
    beta15~dnorm(0.0,0.001);  tau~dgamma(16,6)
for (i in 1 : N) {
     epsilon[i]~dnorm(0.0, tau)
     lambda[i] <-        beta0 +
                         beta1*X9[i]   + beta2*X10[i]  + beta3*X2[i]   +
                         beta4*X3[i]   + beta5*X4[i]   + beta6*X5[i]   +
                         beta7*X6[i]   + beta8*X7[i]   + beta9*X8[i]   +
                         beta10*X11[i] + beta11*X12[i] + beta12*X13[i] +
                         beta13*X14[i] + beta14*X15[i] + beta15*X16[i] +
                         epsilon[i]

    Y[i] ~ dnorm(lambda[i], tau)
        }
}
```

Informed Prior Model with Interaction. The final Bayesian models are specified below. The interaction model is specified immediately below. The interim informed prior model can be obtained by commenting out the interaction term (beta16).

```
Model {
    beta0~dnorm(0.0,0.1);   beta1~dnorm(-.025,0.1);   beta2~dnorm(0.0,0.1)
    beta3~dnorm(0.23,0.1);  beta4~dnorm(0.615,0.1);   beta5~dnorm(-0.068,0.1)
    beta6~dnorm(0.0,0.1);   beta7~dnorm(-.033,0.1);   beta8~dnorm(0.299,0.1)
    beta9~dnorm(0.0,0.1);   beta10~dnorm(0.0,0.1);    beta11~dnorm(0.0,0.1)
    beta12~dnorm(0.0,0.1);  beta13~dnorm(0.0,0.1);    beta14~dnorm(0.0,0.1)
    beta15~dnorm(0.0,0.1);  beta16~dnorm(0.0,0.1);    tau~dgamma(16,6)
```

```
    for (i in 1 : N) {
        epsilon[i]~dnorm(0.0, tau)
        lambda[i] <-        beta0 +
                        beta1*X9[i]    + beta2*X10[i]   + beta3*X2[i]    +
                        beta4*X3[i]    + beta5*X4[i]    + beta6*X5[i]    +
                        beta7*X6[i]    + beta8*X7[i]    + beta9*X8[i]    +
                        beta10*X11[i] + beta11*X12[i] + beta12*X13[i] +
                        beta13*X14[i] + beta14*X15[i] + beta15*X16[i] +
                        beta16*X6*X3   + epsilon[i]

        Y[i] ~ dnorm(lambda[i], tau)
    }
}
```

The model informed by the Chubb and Moe (1990) data can be obtained by replacing the β priors with the following:

```
beta0~dnorm(0.0,0.1);      beta1~dnorm(-.023,0.1)     beta2~dnorm(0.024,0.1)
beta3~dnorm(0.0,0.1);      beta4~dnorm(0.042,0.1);    beta5~dnorm(.0016,0.1)
beta6~dnorm(0.002,0.1);    beta7~dnorm(0.007,0.1);    beta8~dnorm(0.032,0.1)
beta9~dnorm(0.017,0.1);    beta10~dnorm(0.0,0.1);     beta11~dnorm(0.0,0.1)
beta12~dnorm(0.0,0.1);     beta13~dnorm(0.0,0.1);     beta14~dnorm(0.0,0.1)
beta15~dnorm(0.0,0.1);     tau~dgamma(16,6)
```