Bayesian Analytical Methods: A Methodological Prescription for Public Administration

Jeff Gill*, Christopher Witko†
**Washington University; †Saint Louis University*

ABSTRACT

In this article we describe in detail the Bayesian perspective on statistical inference and demonstrate that it provides a more principled approach to modeling public administration data. Because many datasets in public administration are population-level, one-time unique collections, or descriptive of fluid events, the Bayesian reliance on probability as a description of unknown quantities is a superior paradigm than that borrowed from Frequentist methods in the natural sciences where experimentation is routine. Here we provide a thorough, but accessible, introduction to Bayesian methods and then demonstrate our points with data on interest group influence in US state administrative agencies.

INTRODUCTION

This essay introduces Bayesian statistical inference and argues that it provides an ideal research paradigm for empirical scholars in public administration. We work in a field that generally uses data incompatible with standard Frequentist statistical thinking because they are usually population measures that can never be repeated as if in a standard experimental setting. Those working with public administration data need an approach that recognizes that the data are one-time events or collections that usually describe an entire set of objects of interest. Samples also often emerge from survey research collections, and these data are also treated appropriately by the Bayesian approach.

Most nonstatisticians are nonplussed in learning that there are different "philosophical" approaches to statistics. Understandably, basic courses in graduate school emphasize collecting a set of useful tools rather than an introspective look at the theoretical underpinnings of statistical analysis. Unfortunately this leads to practices that are often *wrong* later in one's career. Therefore some research that we read in public administration journals and texts is misleading and unproductive to the discipline,

Our thanks to Cynthia Bowling for sharing and assisting us with the data; Christine Kelleher for pointing us towards data sources; and three anonymous referees for useful comments. Address correspondence to the author at jgill@wustl.edu.

doi:10.1093/jopart/mus091

though it could be easily improved. Surprisingly, little attention has been focused on this problem. One exception, Gill and Meier (2000), argued forcefully for a set of more appropriate practices for data analysis and inference in the field. Yet this tome has fallen on relatively deaf ears, and Bayesian analytical methods remain an apocrypha to empirical scholars in public administration.

Here we seek to improve this state of affairs. Our objective is to describe Bayesian methods and Bayesian statistical inference in detail, but in such a way that it is readable and accessible to a wide audience of public administration scholars. In doing so, we hope to show that the Bayesian perspective is more appropriate for the type of data we encounter in the discipline, provides more intuitive results through probability statements, and is only slightly more difficult to implement than non-Bayesian approaches given the convenience of modern statistical software. To illustrate these points and the Bayesian process, we provide step-by-step tutorial instructions using data from the US states concerning self-reports by state-level administrators about the influence of organized interests on agency budgets and policy. This example is intended to enable new users of Bayesian methods to immediately apply this process to their own data, as well as to show that there is a natural fit between public administration data and Bayesian probability modeling.

There are strong substantive reasons that scholars in public administration should prefer the Bayesian approach to data analysis and inference. Often times the decision-making process for public managers is Bayesian in the sense that executives update past beliefs and actions based on acquired political and administrative information. Boyne, Meier, O'Toole, and Walker (2006, 638), in fact, bluntly state that "managers operate in a Bayesian world." Therefore researchers studying bureaucratic behavior have research objectives which "fits exactly the assumptions of Bayesian statistics and fits poorly with the assumptions of classical statistics" (2006, 638). Many datasets in public administration are population-level, one-time unique collections, or descriptive of fluid events. So the Bayesian reliance on probability as a description of unknown quantities is a superior paradigm than that borrowed from Frequentist methods in the natural sciences where experimentation and detailed control is routine.

In addition, we contribute to substantive debates regarding the influence of organized interests in agency processes. Contracting and loose governance networks have some potential advantages over traditional, top-down, bureaucratic approaches (Kettl 2000), but our findings further highlight some of the downsides of these arrangements. We find that the extensive use of contracting is related to greater perceived interest group influence over agencies (Kelleher and Yackee 2009), and that managers who spend more than a typical amount of time with representatives of organized interests report that those groups have more influence over their budgets and policy. Thus, direct lobbying of public managers results in more influence over agency outcomes, much as in the legislature. We also observe, however, that where legislatures are perceived to be more influential over agencies so too are organized interests. This suggests that rather than constraining interest group influence, close political oversight by the legislature may actually enhance it because legislators often have close relationships with the organized interests attempting to influence the decision making of public managers (Stigler 1971; Witko 2011).

A CLASH OF PHILOSOPHIES

Bayesian statistical inference is based on fundamentally different assumptions about the data and parameters from classical methods. To Bayesians, all quantities are divided into two distinct groups: those that are observed and therefore fixed, and those that are unobserved and must be estimated. These observed quantities are usually the data at hand and values known with certainty. The unobserved quantities are usually the model parameters of interest and any missing data. In the Bayesian world the unobserved quantities are assigned distributional properties and, therefore, become random variables in the analysis. These distributions come in two basic flavors. If the distribution of the unknown quantity is not conditioned on fixed data, it is called *prior distribution* because it describes knowledge prior to seeing data. Alternatively, if the distribution is conditioned on data that we observe, it is clearly updated from the unconditioned state and, therefore, more informed. This distribution is called *posterior distribution*.

This distinction about what is random and what is fixed is at the heart of the debate between Frequentists and Bayesians. Bayesians view observed data as permanently fixed, but unknown parameters are considered random quantities given distributions based on the current level of knowledge. Conversely, Frequentists view data as stochastic, coming from a never-ending stream created by exactly the same generating process, but parameters are quantities fixed by nature and never changing. The latter approach is ideal in industrial quality control or experimental settings in some natural sciences where the researcher has the ability to generate large streams of independent identically distributed (IID) data. This is, unfortunately, not like the data we generally use in public administration research. We typically get a dataset that is situational in time and circumstance and will never be replicated. That is, we cannot go back and re-survey agency executives and assume that no attitudes, experiences, or administrative events have changed. Thus, our datasets represent a fixed, unique look at the phenomenon of interest.

Unfortunately many scholars are confused about Frequentism. This term comes from the canonical work of Neyman and Pearson (1928a, 1928b, 1933a, 1933b, 1936) during the adolescence of statistical science. They posited a small fixed α value established before experiments that eventually becomes the probability of a Type I error after many iterations. Because the data are presumed infinite and IID, it is possible to test very sharp hypotheses, and Frequentist hypothesis tests are always about two possible states of nature that cover all possible states of nature. Therefore *rejection* of one hypothesis means *acceptance* of the other, and neither is labeled as the "null." Wait! We were all (correctly) told in graduate methods courses in social science departments that you can never *accept* a hypothesis because there are an infinite number of alternative hypotheses that have not been tested. What this means is that we do not set up our hypothesis testing and inference engine in a Frequentist fashion. In fact, it would be very difficult to set up a true Frequentist test in public administration since our subjects are not experimentally cooperative and change characteristics over time (even a short period of time).

So, if the bulk of public administration scholars calling themselves Frequentists are really not Frequentists, then what are they? They are actually Fisherian likelihoodists,

although burdened with a poor testing paradigm. The flawed null hypothesis significance test (NHST) is deeply ingrained in the social sciences and it forces an illegitimate (in both senses of the word) blend of Frequentism and Likelihoodism. Fisher's likelihood-based test of hypotheses does not require a priori fixing of an α level and does not require two complementary, exhaustive hypotheses. Instead, it sets up a null hypothesis which is just something to be nullified in Fisher's construct, but is generally used as a claim that there is no specific between-variable relationship in the data. He believed that a prior determination of α was overly rigid and nonscientific. Unfortunately the NHST provides an inconsistent and logically defective combination of these two opposites (Gill 1999).

Fortunately, empirical researchers in public administration do not rigidly adhere to the NHST process and typically just look at evidence from a maximum likelihood–based estimation process to make claims about relationships in data. This is what makes them Likelihoodists. The punchline is this: All likelihood-based models are Bayesian models in which the prior distribution is an appropriately selected uniform prior, and as the size of the data gets large they are identical given any finite appropriate prior. So such empirical researchers are really Bayesian; they just do not know it yet.

With the dramatic increase in Bayesian methods in the social sciences, there are frequent applications to GLMs, causal inference, time-series, change-point problems, ideal point estimation, expert elicitation, missing data imputation, genetics analysis, textual analysis, ecological inference, neural networks, structural equation models, nonparametrics, and factor analysis. This long list is revealing because it demonstrates that Bayesian approaches are not just another "hammer" in the researcher's toolbox but are instead a general philosophical way of thinking about data and estimation. For a recent detailed discussion, see Samaniego (2010).

The Bayesian approach will continue to gain in popularity because it is perfectly suited for the observational data we deal with and the theories we are concerned about in public administration. Almost no scholar in the discipline believes that the phenomenon they care about is fixed and unyielding over time and circumstance. Instead, we care about quantities such as the likelihood that a public policy implementation is effective, the probability of a budgetary outcome, the tendency for administrators to interact with legislators, and so on. These, and other related concerns, are by definition varying quantities and, therefore, best described with distributions as in the Bayesian approach. Also, public administration is a field tied to history in the sense that, unlike the other social sciences, we can point to a record of programs, behaviors, and policies in government that have been assessed by scholars, journalists, and other observers over time. Thus a research paradigm that lets previous information, both qualitative and quantitative, to be systematically included into the modeling process is ideal for studying the administrative working of government.

BAYESIAN FOUNDATIONS

The Bayesian inference process specifies prior distributions for unknown parameters and updates these to become posterior distributions using observed data contained in

the standard likelihood function. So this process is really just a slight variant of conventional likelihood inference, which includes previously known relevant facts. This updating step can be repeated as new data are observed, and therefore the Bayesian inference process is a principled incremental process of scientific discovery. In this section, we carefully step through the stages of Bayesian inference in a social science setting (for additional details, see book-length expositions such as Gill 2008 or Gelman, Carlin, Stern, and Rubin 2003). Our exposition starts with an arbitrary data vector, X, whose distribution is conditioned on an unknown parameter, β, and then proceeds to the standard regression setting where a matrix of explanatory variables, $\mathbf{X}$, affects an outcome variable vector, $\mathbf{y}$, conditional on an unknown parameter vector, β.

The Prior Distribution

Prior distributions range from very informative descriptions of previous research in the field to purposefully vague forms that reflect relatively low levels of prior knowledge about the effect in question. This level of information is dependent on the volume and reliability of previous studies on the topic of interest. This part of the process is not a necessary inconvenience imposed by the process. Instead it is an opportunity to systematically include qualitative, narrative, and intuitive knowledge into our statistical models. There is a lot of historical controversy surrounding the assignment of prior distributions, and this leads many applied Bayesians in the social sciences to use highly diffuse forms such as the uniform distribution. But informed prior distributions are incredibly useful for integrating nonquantitative information into the statistical model. The two critical requirements for informed prior distributions are defending the source of the information used and showing the impact of this prior distribution relative to some reference form.

Prior distributions are probability statements about some parameter of interest, β, that are not conditioned on the data being considered, denoted $p(\beta)$. We could stipulate a uniform distribution by using $p(\beta) = k, a < \beta < b$, meaning that there is constant probability that β takes on some value between a and b. If we assert that $\beta \sim \mathcal{N}(\mu, \sigma^2)$, then this is a belief that before conditioning on new data β is normally distributed around μ with variance σ^2. Frequently, if there is little prior information about this parameter, then researchers can use a normal distribution with $\mu = 0$ and large σ^2, in order to give a vague or cynical statement about the efficacy of β as an important phenomenon. It is critical, also, to give a prior distribution that has the same support (defined over the measured support on the x-axis) as the parameter in the model. For instance, in modeling a variance components parameter, we would want to stipulate a prior distribution bounded by $(0, \infty)$. This makes the gamma distribution a popular choice for Bayesian specifications of variance terms.

Where do prior distributions come from? Often there is substantial guidance from previous studies by the researcher or strong suggestions from published work. In the latter case, a meta-analysis of some germane literature may strongly suggest parameters for a normal or students-*t* form. There is also a literature on "expert elicitation" whereby substantive experts, usually with no interest in or connection to the statistical aspect of the study, are systematically queried in such a way that prior distributions are

produced from their qualitative responses (Gill and Freeman 2013; Gill and Walker 2005). Prior elicitation is often straightforward for obtaining prior means but challenging for getting an informed view of the prior variance from elicitees as uncertainty can be more difficult to translate. We might also want to recognize that researchers with deep contextual knowledge, and a long history of studying some question in public administration, may have enough intuition to specify a personalized prior, even if this requires substantial justification to readers. A very popular approach is to specify *conjugate priors*, which is a mathematically convenient form whereby the distributional family of this prior flows through to the final distribution (albeit with different parameterization). See the developed example on page 21 where the prior distribution for the linear regression parameters is specified as a normal (Gaussian) form and the model produces normal distributions for the final parameter estimates. Sometimes it is reasonable to use other data sources or a fraction of the current data to empirically produce a prior distribution in the same way that histogram or smoother implies a distribution, although this can be controversial. Finally, many authors specify priors for diagnostic purposes, such as using a uniform distribution to show a contrast with some more informed version from one of the sources just described. This can show the relative "influence" of some desired form relative to a benchmark that readers easily identify with. The key point is that priors can come from many sources, and as long as the justification is reasonable then the resulting specification is principled. It is also important to observe that the overwhelming proportion of prior distributions specified in published Bayesian social science work still avoids using reasonably informed priors, which unfortunately hurts the steady accumulation and progression of scientific knowledge. Note finally that statistically *unreasonable* priors are: those that are not specified on the same support (the range of the variable described), forms that have high density in substantively illogical regions, and expressions that are unnecessarily mathematically complicated given the phenomenon being described. See Gill (2008, chapter 5) for a detailed discussion of relevant issues.

The Likelihood Function

The second step is to stipulate a likelihood function in the usual manner by assigning a parametric form for the data and plugging in the observed data. This step is done in *exactly* the same manner as any other likelihood-based statistical model, as commonly practiced in public administration. Note that this process is just as subjective as the choice of the prior distribution. This means that the researcher needs to justify this parametric form to readers as well.

Some collected data $\mathbf{X}$, an $n \times 1$ vector of observations, are treated Bayesianly as a fixed quantity and we make a reasoned assumption about the probability mass function (PMF) or probability density function (PDF) for describing the original data generation process conditioned on a single unknown parameter, β. The well-known maximum likelihood estimation process substitutes the unbounded notion of likelihood for the bounded definition of probability by first considering a function that is the joint distribution of the observed data:

$$p(\mathbf{X}\,|\,\beta) = p(\mathbf{X}_1\,|\,\beta)p(\mathbf{X}_2\,|\,\beta)\cdots p(\mathbf{X}_{n-1}\,|\,\beta)p(\mathbf{X}_n\,|\,\beta), \tag{1}$$

where $p(\mathbf{X}_i \mid \beta)$ is the assigned PMF or PDF. Fisher's (1922, 1925a, 1925b, 1934) notion was to turn this around logically, noting β to be the unknown value and $\mathbf{X}$ to be the known values. Accordingly, equation (1) is relabeled as $L(\beta \mid \mathbf{X})$, which is of course the likelihood function. Note that this is subtly a very Bayesian process in the way described above since the data are now the fixed quantity.

The process continues in finding the value of β that is most "likely" to have generated the observed data. This is a simple process since the standard forms of the PMF and PDF used are guaranteed to produce a unimodal function concave to the x-axis. Therefore there is a unique value at the top of this hill, $\hat{\beta}$, that maximizes the function $L(\beta \mid \mathbf{X})$. Furthermore, this value is easily found with standard calculus tools: Take the derivative of $L(\beta \mid \mathbf{X})$ (called the score function) to produce a function that gives the slope of the tangent line for any give β, then figure out where this tangent line has slope zero, since that is the unique point at the top of the unimodal function. The second derivative of the likelihood function at this modal point leads to the variance of the estimate (through the negative inverse expected value). Bayesian inference actually goes further than this conventional process because it estimates and describes the full distribution of the results, not just a point estimate and curvature around it.

The Posterior Distribution

The third step is to produce a posterior distribution by multiplying the prior distribution and the likelihood function, which is the numerator of Bayes' Law: $p(A \mid B) = p(B \mid A)p(A)/p(B)$. So the likelihood function uses the data through the likelihood function to *update* prior knowledge into posterior knowledge. The posterior distribution represents the most informed set of knowledge about the phenomenon of interest because it is the most updated version available. It is also important to note that the posterior distribution tells us *everything we know* about the effect of interest through a distribution. This is more information than the typical summary from a likelihood analysis: a single point estimate and a measure of curvature around it. A distributional summary allows more description: the mode, mean, or median of the distribution, as well as any quantiles of interest.

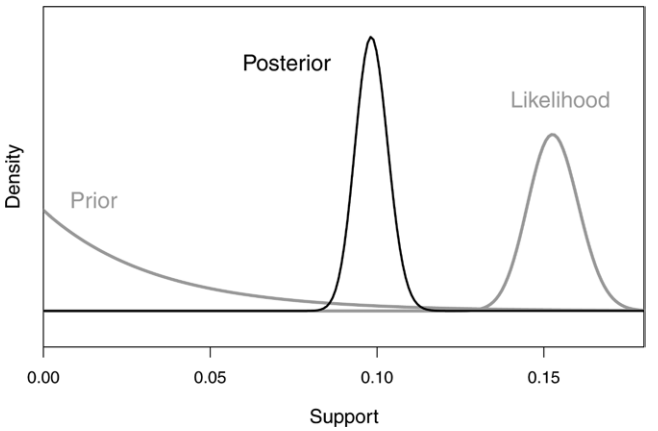
Suppose we have a single parameter of interest, denoted β, and a vector of data, denoted $\mathbf{X}$. Then the steps of Bayesian inference are summarized in the proportional version of Bayes' Law:

$$\text{Posterior probability} \propto \text{Prior probability} \times \text{Likelihood function}$$

$$\pi(\beta \mid \mathbf{X}) \propto p(\beta) \times L(\beta \mid \mathbf{X}), \tag{2}$$

where $\pi(\beta \mid \mathbf{X})$ is the resulting posterior distribution from conditioning the prior distribution, $p(\beta)$, on the likelihood function, $L(\beta \mid \mathbf{X})$. Here $\pi()$ is used to distinguish the posterior distribution from the prior distribution, $p()$. The use of proportionality ($\propto$) may seem curious here. This results from leaving out $p(\mathbf{X})$ from the denominator of equation (2). Recall that to a Bayesian the data are fixed once observed, meaning that the quantity $p(\mathbf{X})$ actually has no probabilistic properties to account for in the calculation of $\pi(\beta \mid \mathbf{X})$ here. Therefore the only purpose for this quantity in the

Figure 1
Posterior Production



calculation of the posterior distribution is to make the posterior distribution fully normalized: summing or integrating to one over its total support. Since this is an easy value to recover later in the process,[1] it is usually more convenient to ignore this difference in the calculation and to renormalize at the final step. As we shall see on later, $p(\mathbf{X})$ is useful for model comparison purposes.

The production of a posterior distribution is described in figure 1. Here we show that the posterior distribution is a compromise between prior information and likelihood information. When the data size is large the likelihood function has more influence and pulls the posterior closer to its location. Conversely, when the data size is small and the prior distribution has a nondiffuse form, the resulting posterior is closer to this prior. These countervailing effects are called *shrinkage* and we can measure how much the likelihood function shrinks some statistic, say the posterior mode, towards its mode, away from the prior distribution.

The posterior distribution can also be treated as a new prior distribution if additional data are later observed. In this way the parameter of interest is updated and knowledge is accumulated over time. Suppose $\pi_1(\beta \mid \mathbf{X}_1)$ is the posterior from equation (2) with the data $\mathbf{X}_1$. Later a new dataset, $\mathbf{X}_2$, is observed and we treat $\pi_1(\beta \mid \mathbf{X}_1)$ as a prior for a second update. Then $\pi_2(\beta \mid \mathbf{X}_2) \propto \pi_1(\beta \mid \mathbf{X}_1) L(\beta \mid \mathbf{X}_2)$ is a new posterior distribution. Interestingly this is the same distribution we would get if $\mathbf{X}_1$ and $\mathbf{X}_2$ arrived together and we created a posterior based on them at once, $\pi_2(\beta \mid \mathbf{X}_2) = \pi_{1,2}(\beta \mid \mathbf{X}_1, \mathbf{X}_2)$.

Summarizing Bayesian Results

Unlike the seriously flawed NHST, evidence from a Bayesian model is presented by summarizing the posterior distribution in various informative ways. This is typically

1 Suppose we end up with a nonnormalized posterior distribution for β that sums or integrates to something other than one, say 2.3 just to make up a value. Then simply dividing the PDF or PMF by 2.3 would return the probability statement to a regular normalized distribution in every sense.

done with distributional quantiles and probability statements. These are interpreted as the probability that the parameter of interest is less than/greater than some constant, or the probability that this parameter occupies some region of the support, for example $p(\beta > 0)$. It is useful to be able to say something like the (posterior) probability that previous government experience for an agency head could lead to a better working relationship with the legislature is 0.93. Notice that such statements are a function of the posterior mean and the posterior variance, both of which are determined jointly by the prior distribution and the likelihood function. Therefore we care not only about where the posterior is centered but also how dispersed it is around this point, which indicates relative uncertainty. Such posterior statements are also free of an explicit null hypothesis or a null statement, although the zero point is often implied as one. This means that we are free from the encumbrance of NHST fixations and can make statements of substantive interest to public administration scholars in simple probabilistic terms that focus on the question of interest.

There is no reason that the Bayesian posterior distribution cannot be simply described with a mean and standard deviation (which corresponds to a standard error in conventional models). This means that authors can build a regression table that looks like regular forms, except of course that "stars/asterisks" are inappropriate. So Bayesian results can be given in very accessible formats and readers do not need to learn some new exotic reporting criteria.

Interval Summaries

There is also a Bayesian analog to confidence intervals called the *credible interval*. This is constructed in exactly the same way as a confidence interval (a point estimate plus/minus some critical value times the standard error). However, the interpretation is different and it is exactly the interpretation that first-time readers mistakenly assign to the confidence interval: the probability that some effect exists between two bounds. The Frequentist confidence interval is really an interval that over $1/\alpha$ replications of the exact same experiment cover the true fixed value of the parameter on average with probability $1-\alpha$.

There is also a more flexible version of the Bayesian confidence interval, called the highest posterior density (HPD) interval. This is the region of the support of β that contains the highest $1-\alpha$ posterior density regardless of whether it is continuous or not. So the HPD interval can be multiple intervals with multimodal posterior forms. Like Frequentist confidence intervals, an HPD region 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 over identical replications, an HPD interval provides a $(1-\alpha)\%$ probability that the true effect is in the interval.

Testing

Hypothesis testing with an explicit decision can also be performed in the Bayesian setup. Bayesian hypothesis testing is less fixated with a "null" model, although it is common to evaluate estimated regression effects relative to the zero point. More commonly

Bayesians seek to compare two plausible models against each other (Raftery 1995). Suppose β_1 and β_2 represent two competing hypotheses about the state of β. These two form a partition of the sample space: $\mathcal{B} = \beta_1 \cup \beta_2$, and $\beta_1 \cap \beta_2 = \varnothing$. First prior probabilities are assigned to the two outcomes: $\pi_1 = p(\beta \in \beta_1)$, $\pi_2 = p(\beta \in \beta_2)$. Therefore there are two resulting posterior distributions from the application of two different priors on the same likelihood function: $p_1 = p(\beta \in \beta_1 \mid \mathbf{X})$ and $p_2 = p(\beta \in \beta_2 \mid \mathbf{X})$. Now we just need to compare these posteriors to test H_1 versus H_2.

The prior odds are defined by the ratio p_1/p_2, and the posterior odds, π_1/π_2. These can be combined as the ratio of ratios $(\pi_1/\pi_2)/(p_1/p_2)$ (posterior odds over prior odds), which is called the Bayes Factor (Kass and Raftery 1995). The Bayes Factor is interpreted as odds favoring H_1 versus H_2 given the observed data. So if the Bayes Factor is much larger than one the test favors H_1, and if the Bayes Factor is much smaller than one the test favors H_2. Values around one indicate a lack of evidence favoring either hypothesis. This is one area where the omitted denominator of equation (2) is important. To provide a more complete version of Bayes' Law than that using proportionality, we would include the denominator $p(\mathbf{X}) = \int_{\beta} p(\beta) L(\beta \mid \mathbf{X}) d\beta$. This term is typically called the *marginal likelihood*, 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 $\pi(\beta \mid \mathbf{X})$ sums or integrates to one in the expression of Bayes' Law, which for models (e.g., hypotheses) means that:

$$\pi(M \mid \mathbf{X}) = p(M)/p(\mathbf{X}) \times p(\mathbf{X} \mid M), \tag{3}$$

where $p(M)$ denotes the prior on the model (hypothesis).

The intuition behind Bayes Factors in a regression setting is best understood with the abstraction of two competing models, M_1 and M_2, defined by two sets of possible sets of explanatory variables, β_1 and β_2, using the same data $\mathbf{X}$. Interestingly, β_1 and β_2 do not need to be nested as with a standard likelihood ratio test, which is a major advantage over non-Bayesian approaches. The prior vectors for these regression parameters are given by: $p_1(\beta_1)$ and $p_2(\beta_2)$, and since the better model is also an unknown we can assign prior probabilities: $p(M_1)$ and $p(M_2)$. The posterior odds ratio in favor of Model 1 against Model 2 is produced by Bayes' Law:

$$\underbrace{\frac{\pi(M_1 \mid \mathbf{X})}{\pi(M_2 \mid \mathbf{X})}}_{\text{posterior odds}} = \underbrace{\frac{p(M_1)/p(\mathbf{X})}{p(M_2)/p(\mathbf{X})}}_{\text{prior odds/data}} \times \underbrace{\frac{\int_{\beta_1} f_1(\mathbf{X} \mid \beta_1) p_1(\beta_1) d\beta_1}{\int_{\beta_2} f_2(\mathbf{X} \mid \beta_2) p_2(\beta_2) d\beta_2}}_{\text{Bayes Factor}}. \tag{4}$$

So the quantity of interest turns out to be the ratio of marginal likelihoods from the two regression models. Unfortunately the Bayes Factor can be difficult to calculate numerically with elaborate regression models. This is still an elegant testing structure that clearly shows the advantage of one model over another is simple settings (if there is one). However in more elaborate regression settings, the calculation of equation (4) can be challenging. Therefore applied Bayesians have sought more computationally convenient comparison tools, the most important of which we now describe.

A more modern testing tool is the Deviance Information Criterion (DIC; Spiegelhalter et al. 2002). This measure, commonly called the DIC, is designed to

be a Bayesian version of the well-known AIC, the Akaike Information Criterion (Akaike 1976), which was designed to balance model fit and covariate parsimony by comparing the log-likelihood for a fit model to the number of parameters (p) used: $\text{AIC} = -2\ell(\hat{\beta} \mid \mathbf{X}) + 2p$. Consider again a model likelihood or posterior defined by $p(\mathbf{X} \mid \beta)$ for data matrix $\mathbf{X}$ and "true" parameter vector β. The "Bayesian Deviance" for this model is minus two times the log of this quantity plus two times the marginal likelihood of the data only, which unwinds the ratio of the likelihood and the data component in log terms.

$$D(\beta) = -2\log[p(\mathbf{X} \mid \beta)] + 2\log[f(\mathbf{X})], \quad (5)$$

and if we substitute in an estimate of β then this expression can be thought of as the *deviance of the means*. It turns out that the $2\log[f(\mathbf{X})]$ term is unimportant since it will cancel out in model comparisons, as we will see. So now modify this expression by averaging (taking expectations) over the parameters according to this distribution to get a *mean deviance* (difference) measure of model fit:

$$\bar{D} = E_{\beta}[-2\log(p(\mathbf{X} \mid \beta)] + 2\log[f(\mathbf{X})] = -2\int_{\beta} \log[p(\mathbf{X} \mid \beta)]d\beta + 2\log[f(\mathbf{X})], \quad (6)$$

Intuitively, $\int_{\beta} \log[p(\mathbf{X} \mid \beta)]d\beta$ is smaller for better fitting models because the distribution $p(\mathbf{X} \mid \beta)$ is closer to the actual underlying phenomenon that we are trying to estimate with the model. This expectation (integration) sounds hard but we basically get its calculation "for free" through Markov chain Monte Carlo (MCMC) estimation described in the next section since it generates samples for a numerical version of this calculation automatically as part of the β estimation process. Now notate $\tilde{\beta}$ as a point estimate of the coefficient vector β. This is a flexible definition but in practice it is usually the posterior distribution mean. The difference between $\bar{D}$ (which integrates over the parameter space) and substituting $\tilde{\beta}$ into equation (5) gives a single value:

$$p_D = \bar{D} - D(\tilde{\beta}) \quad (7)$$

This is, therefore, the "mean deviance minus the deviance of the means," and is interpreted as the *effective dimension of the model*, which is analogous to the number of parameters p, in the AIC. The p_D has to be a more nuanced complexity measure since parameters in a Bayesian hierarchical model have varying roles depending on their placement in the levels of the model (e.g., dependencies on other parameters) and the restrictions placed by different prior specifications.

The DIC is the difference between model fit and model dimension,

$$\text{DIC} = \bar{D} + p_D = 2\bar{D} - D(\tilde{\beta}), \quad (8)$$

giving a trade off in the AIC sense, except that the AIC uses a single plug-in value for the fit component rather than integrating over unknown (to be estimated) parameters. So in comparing two models we can claim that the one with the lower DIC value provides the better compromise between model fit and covariate parsimony. Conveniently, commonly used Bayesian estimation software provides this value for a specified model and we are freed from the mechanics of the calculations above.

EVERYBODY IS A BAYESIAN

As different as the Bayesian inferential process may initially appear, we can see that it subsumes likelihood analysis. We have already noted that scholars in public administration do not generally have the ability to be Frequentists due to the nature of the objects we study. So the real comparison of interest is between the standard likelihood-based approach and the Bayesian approach. The maximum likelihood estimate is equal to the Bayesian posterior model with the appropriate uniform prior, and they are asymptotically equal given *any* prior: Both are normally (Gaussian) distributed as the data size gets large. Furthermore, in many cases the choice of a prior is not especially important since as the sample size increases, the likelihood progressively dominates the prior for any reasonable choice of the prior.

Although the assignment of a prior distribution for unknown parameters can be seen as subjective (actually *all* statistical models are subjective), 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 (conjugacy); and the prior allows researchers to include additional information collected outside the current study (as we do in the examplein a later section).

BAYESIAN STOCHASTIC SIMULATION

The post-1990 Bayesian estimation engine of MCMC is the most powerful vehicle for obtaining model results available in statistics. MCMC was introduced into the general statistical literature by Gelfand and Smith in a 1990 review article that appeared in the *Journal of the American Statistical Association* after the idea was lying undetected in statistical physics for almost 40 years. Bayesian stochastic simulation, which is a descriptor of MCMC, replaces pen-and-paper analytical calculations and standard software-driven numerical mode finding with an iterative computational process that explores and describes multidimensional posterior distributions, which may be impossible to integrate. Integration is necessary here to go from a joint probability statement of many dimensions to a regression table that describes each coefficient effect in marginal (individual) terms.

This MCMC process is done by running a *Markov chain*, which will wander through the sample space preferring high-density areas in proportion to the underlying posterior probabilities. Consider the posterior distribution of interest as a geographic region to be described, say Central Park in New York City. Our Markov chain is nothing more than a robot that walks randomly around the park, storing the elevation values where it visits on its hard drive. The Markovian process is the probabilistic component that dictates moves of the robot from one place in Central Park (the sample space) to another. Furthermore, the structure that underlies these probability decisions is conditional only on where the robot is right at that moment: It does not care where it has visited before. Conversely, a maximum likelihood estimation robot would go to the highest elevation in Central Park and never leave.

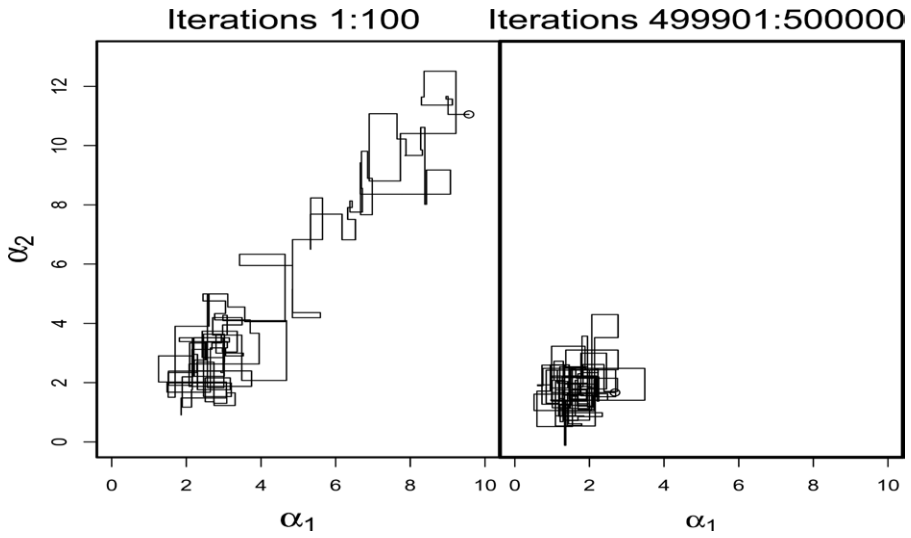
Since each step of the chain is a multidimensional position, marginalizing the joint posterior is simply equivalent to looking at the history of each dimension individually. In the case of our robot we would look separately at the latitude or the longitude values at each step of its walk, ignoring the other, and that is all is necessary for marginalization. Marginalizing is what we want since a row of the regression table is

just a marginal summary of a particular coefficient estimate. Generally this process is straightforward with modern software, as described in a later section. Some challenges include assessing convergence of the chain, getting efficient mixing through the sample space, and setting up the initial probability statements (Gill 2008). Convergence assessment is a mechanical process that uses simple diagnostics to describe stability of the Markov chain at any moment in time with the objective of claiming that the chain is in its stable (stationary) distribution that describes the posterior of interest. We walk through these steps in detail for our empirical example in a later section and supply theoretical details in Appendix B.

There are two principle MCMC algorithms. Gibbs sampling draws iteratively through a complete set of conditional probability statements for each estimated parameter in the model. So the user, or the software, needs to express probability functions where each parameter is on the left side of the conditional statement and any parameters plus the data are on the right side of the conditional. For parameter β_k and a parameter vector not including this parameter, β_{-k}, this looks like $\pi(\beta_k \mid \beta_{-k}, \mathbf{X})$. The Metropolis–Hastings algorithm performs a single multidimensional move on each step by drawing from a proposal distribution and making an accept/reject decision based on the quality of this draw. If the new point has higher posterior probability than the current position of the Markov chain, then the decision is always to move to that point. If it has lower posterior probability than the current point, then the Markov chain will move there based on the ratio of probabilities: $\pi(\beta_{\text{candidate position}})/\pi(\beta_{\text{current position}})$. Therefore unlike the Gibbs sampler, the Metropolis–Hastings algorithm may reject moving to a new position and choose to make the current position the next step in the Markov chain.

To more fully demonstrate the MCMC process, consider figure 2 where we follow the path of the first two parameters from the model estimated below with a Gibbs

Figure 2
Gibbs Sampler Paths for Two Selected Parameters



sampler. Since the posterior distribution being explored here is 76 dimensions (75 parameters in the model plus posterior density, as described in the next section), we cannot show the full structure. However, picking two dimensions gives a view of the behavior of the Markov chain that can be extrapolated to other dimensions. The first panel shows the path of the first 100 iterations of the Markov chain. Notice that the first draw of the chain is in the upper right hand side of the parameter space at [9.5695,11.0518] as indicated by the circle after being started at [10,10]. The chain then works progressively towards a region roughly centered at 2.5 on both axes. It clearly prefers to be in this area since it is near the high-density region of the bivariate density shared between α_1 and α_2. It is important to observe that the Markov chain has not yet converged to its stationary (permanent) distribution at this point. This is why it is important to "burn-in" the Markov chain in practice, meaning run the chain for some lengthy period and dispose of these iterations before beginning to record chain visits for inferential purposes. More discussion of burn-in and convergence is provided in Appendix B. In the second panel we show the last 100 iterations of the chain that is run for half a million iterations. Since the panels are on the same scale it is easy to see that the latter part of the Markov chain is much more stable and much more concentrated. This is because the chain is now almost certainly in its stationary distribution. These recorded values are, therefore, reliable draws from the bivariate posterior of interest for inferential purposes. Notice also that the Markov chain only makes orthogonal moves (90-degree turns or straight moves). This is due to the separate draws of the full conditional distributions in the Gibbs sampler: Draws are done one at a time through a mechanical iterative process that produces, in this example, α_1, then α_2, then α_1 again, the α_2 again, and so on.

Although this may sound quite involved from a programming perspective, there exists extremely high-quality *free* software that considerably lessens the burden on users. Both `JAGS` ("Just Another Gibbs Sampler") and `WinBUGS` ("Windows Bayesian inference Using Gibbs Sampling") allow the user to simply make modeling statements and then let the software convert these into the full conditional distributions needed for Gibbs sampling (the default engine for both). Furthermore, these packages will automatically switch to Metropolis–Hastings for any parameters where this Gibbs process is problematic.

Importantly, MCMC, either Gibbs sampling or the Metropolis–Hastings algorithm, is actually *more* powerful than maximum likelihood; we just do not know it yet. It was actually about 40 years from Fishers' important MLE papers until the full set of properties were described by Birnbaum (1962). The reasons are clear why MCMC is more powerful: It gives the same information as MLE (the mode and curvature around the mode), it gives full information about the posterior distribution so that quantities like quantiles and Bayes Factors can be determined, and the *process* can reveal information on the way (especially in Bayesian nonparametrics, see Gill and Casella 2009). We provide additional theoretical details on Bayesian stochastic simulation in Appendix B and the exact code to run the empirical example in Appendix C. Furthermore, there are estimation demands from models that are increasingly being used in public administration such as structural equations models to capture complex substantive relationships, measurement models for latent variables (e.g.,

Item-Response Modeling), text analysis and mining with latent Dirichlet allocation, and others. Standard maximum likelihood estimation approaches either struggle to provide estimates or simply cannot provide estimates for these settings with realistically large and complex data due to high dimensionality, complex functional forms, or identifiability, whereas MCMC approaches work exceptionally well in such situations.

EMPIRICAL EXAMPLE: CONTRACTING AND INTEREST GROUP INFLUENCE IN STATE AGENCIES

Here we consider an important question in public administration, describe a relevant dataset, and then walk through the Bayesian modeling steps in a systematic, mechanical fashion, with the individual steps numbered for easy identification. Our stated objective is to provide an easy-to-adopt template for constructing and estimating Bayesian hierarchical models in the context of empirical public administration research.

We demonstrate the utility of the Bayesian approach by incorporating prior information from previous studies into an analysis of interest group influence in state administrative agencies using new data. How external political principals influence government agencies is a core public administration question motivating numerous studies (Kelleher and Yackee 2006; Meier and O'Toole 2006; Miller 1987; Miller and Wright 2009; Moe 1985; Potoski 1999; Waterman and Meier 1998; Waterman, Rouse and Wright 1998; Wood and Waterman 1991). In recent years, stories about "special interests," from banks to oil companies and military contractors, using their political influence to undermine the ability of agencies to effectively implement policies have been a media staple. Yet compared to studies of how the elected branches of government influence agencies, fewer analyses directly examine the influence of organized interests in the bureaucracy.

Few organized interests have a greater stake in bureaucratic deliberations than those entities that contract with agencies to provide goods or services (Witko 2011). "Contracting out" is often advocated on efficiency grounds, but Kelleher and Yackee (2009) argue that the relationships established during contracting processes may also increase the influence of organized interests over agency deliberations by ensuring that some organized interests—contractors—will have easy access to public managers. They find evidence for this using data from the 1998 and 2004 rounds of the American State Administrator's Project (ASAP) survey.

If contracting increases group influence this is problematic since the businesses that often contract with government are already over-represented in the bureaucracy (Reenock and Gerber 2008; Yackee 2006). Although the argument is intriguing, this potential negative aspect of contracting, like the interactions between organized interests and the bureaucracy more generally, has not been studied often.

Therefore we consider how contracting may shape interest group influence incorporating the results from Kelleher and Yackee's (2009) study *as prior distributions* for a new round of the ASAP survey data from 2008 (they used 1998 and 2004 iterations of the survey). We also created prior distributions from the information in an important study (Brudney and Hebert 1987) that used the 1978 round of ASAP data to

examine the influence of interest groups (and other external actors) in agencies, using an outcome variable very similar to that used by Kelleher and Yackee (2009). As we will show, the Bayesian approach allows us to easily and explicitly incorporate such information from previous studies into new models.

Data Description

ASAP data have been used to examine how external actors influence government agencies (Brudney, Fernandez, Ryu, and Wright 2005; Brudney and Hebert 1987; Kelleher and Yackee 2006) and many other topics over the decades since it was first conducted in 1964 (e.g., Bowling and Wright 1998; Bowling, Jones, and Kelleher 2006).[2] This study asks a large number of questions of hundreds of state administrators from each of the 50 states over many years and is, therefore, the most comprehensive survey of state administrators in existence. For the purposes of this illustration, we will not consider the issue of common source bias. The ASAP survey asks administrators about the influence of a variety of external political actors including "clientele groups" in their agencies. Organized interest or interest group is a more contemporary term than "clientele group" but it is reasonable to use the terms interchangeably as previous studies have done (Brudney and Hebert 1987; Kelleher and Yackee 2009). Thus, following from Kelleher and Yackee (2006, 2009) and Brudney and Hebert (1987), our outcome variable of interest, `grp.influence`, is an index of the respondents' (senior executives) perceptions of the influence that clientele groups have on the total agency budget, specialized program budgets, and agency policies. Each of these questions is a seven-point scale and they have been summed to create a single outcome variable (ranging from 3 to 21).

We include the same explanatory variables as Kelleher and Yackee's (2009), with a couple of exceptions noted below. To measure the level of contracting performed by the respondent's agency, we include the variable `contracting`, which is coded from zero to six, where increasing values indicate higher levels. Because more time spent with groups may increase their influence, we also control for whether the administrator in question spent more than the median amount of time with interest group representatives, `med.time`, which is slightly different from Kelleher and Yackee's (2009) who used a dummy variable indicating whether the amount of time spent with groups was above the average.[3]

In addition, we include a variable used in the Brudney and Hebert (1987) study which significantly shaped group influence, the method of appointment of the agency head. Brudney and Hebert (1987) found that elected officials or those appointed by a board or commission were most likely to be influenced by interest groups. This reflects that these groups play an important role in elections and can wield significant influence in the selection processes of relatively obscure boards and commissions, because a lack of visibility generally enhances interest group influence (Witko 2006). Therefore we control for whether the respondent is an elected official or appointed by a board or commission, rather than a merit employee or appointed by the legislature and/or

2 For a more thorough list of publications through the mid-2000s, see http://www.auburn.edu/outreach/cgs/documents/(091007)ASAPinventoryofpublicationsandpapers.pdf.

3 In the 2008 data, very few groups spent above the average amount of time; so we use the median.

executive (`elect.board`, 1 if so, 0 otherwise). Using this variable contrasts with Kelleher and Yackee (2009) who controlled for whether the agency head was a civil servant (which was not a significant predictor of group influence). Following from Kelleher and Yackee (2009), we include the multiplicative interaction between time and contracting `medt.contr`, which is anticipated to increase group influence.

It is also necessary to control for several factors that may enhance or limit interest group influence. We followed Kelleher and Yackee (2009) in selecting these controls. First, since elected political principals may condition interest group influence by imposing their own will on the bureaucracy, we include two control variables asking about the influence of the governor and legislature constructed in the same manner as the interest group influence outcome variable, `gov.influence` (3 to 21) and `leg.influence` (3 to 21). We also include individual-level administrator variables for the number of years in the current position (`years.tenure` from 0.25 to 40), level of education (`education` from 1 to 5), indicating high school, some college, bachelors degree, graduate study, and graduate degree), and partisan identification (`party.ID` 0 to 5, moving from strong Democrat to strong Republican). Next we include a set of 11 dichotomous explanatory variables that control for the type of agency using Deil Wright's 13 functional categories (thus adding a net of 10 columns to the **X** matrix). There are 11 dummy variables because we omit one category ("other"— i.e., agencies that do not fit in one of the standard categories) as the reference group and we also omit the category for agencies headed by elected officials since this is perfectly determined by our variable of theoretical interest `elect.board` derived from the Brudney and Hebert (1987) study. The actual definitions of the 11 agency types are transformed into distinguishable but unlabeled categories to protect the respondents' anonymity as a condition of the use of the data. This is slightly inconvenient but important because it would otherwise be possible, although very difficult, to reconstruct personal identities from the data plus other publicly available references. Furthermore, we also exclude gender from this analysis since the low number of females makes that group vulnerable to discovery, and we add $N(0,1/3)$ random noise to `years.tenure` to obfuscate this variable without adding bias. Such data manipulation processes have a long history in public administration and is one of the only ways to get public figures to answer sensitive political questions in an honest and forthright manner (Mackenzie and Light 1987).

As in Kelleher and Yackee (2009), at the state level of the hierarchical model (described below) we specify three explanatory variables. First, `gov.ideology` is an omnibus measurement of the state-level government ideology (Berry, Ringquist, Fording, and Hanson's, 1998, updated measure). The volume of organizations seeking to influence government may matter, so we include `lobbying` (ranging from 191 to 2126), which is the total lobbying registrants in 2000–1 from Gray and Lowery (1996, 2001). Finally, we use the variable `nonprofits` to count a balancing force: the number of nonprofit organizations registered in the states as of 2008 (originally ranging from 4,537 to 156,682 and therefore scaled by 10,000) from the National Center for Charitable Statistics. Notice that these definitions at the state level ("group level" in multilevel modeling language) make them inappropriate for assignment to the individual administrators answering the survey since these people are nested in states with other public managers, not state-defining units themselves. Ignoring such aggregation in the data with a "fully pooled" or flat model that does not incorporate this hierarchy leads to incorrect model results, such as inappropriately small coefficient standard errors (Gelman and Hill 2007).

One departure from the Kelleher and Yackee (2009) model was that we did not include a variable measuring the influence of professional associations over the agency as a control in the model, because we think that professional associations could also be viewed as a type of group seeking to influence the agency (either where they represent employees working in agencies, or are even regulated by agencies in some cases), and the two are indeed relatively highly correlated in our data ($r = 0.61$). We did not think it was proper to incorporate professional association influence into the outcome variable, however, because professional associations rarely contract with the state government. In addition, since we used one round of data in our analysis we do not include a dummy variable for the different rounds of data.

Our version of the raw 2008 ASAP data with 713 observations had 933 missing values within some variables, giving 2.75% total missingness (933 divided by 713 observations times 47 variables). Footnote 12 in the original article suggests that Kelleher and Yackee (2009, 595) case-wise deleted full observations with missing values in the 1998 and 2004 ASAP dataset before running their regression models. Since this practice typically leads to biased model specifications (Little and Rubin 1983, 2002; Rubin 1987), we imputed the missing values with the mice multiple imputation package in R and the tools for discrete missing values given in Cranmer and Gill (2013). This process gives us a set of 10 fully filled-in datasets (of size 713 observations times 47 variables) that require replicated modeling and averaging of inferential results (there is a slightly more involved process for combining the standard errors). See the review essay by Rubin (1996) or the text by Schafer (1997) for details. It is also possible to draw values of these missing data during the iterative Gibbs sampling process, but pre-estimation multiple imputation is often easier when using the packages `JAGS` and `WinBUGS`. These packages treat `NA` symbols in data statements as missing values, and in some situations can estimate them in the same way as parameters in the model estimation process since everything unknown to a Bayesian is treated distributionally, although both packages are restrictive in their implementations.

Model Development

This section outlines the model specification process. Here we build and estimate a Bayesian version of the specification in Kelleher and Yackee (2009) with the noted enhancements designed to reveal more information about the data generating process. In addition, we provide the detailed data handling and estimation code for processing this model in `JAGS`. The program `WinBUGS` is very similar but we focus on `JAGS` since there is evidence that it is a better MCMC engine, *and* it runs on all common operating systems. There are two primary ways to run `JAGS`: natively in terminal window where the results are imported into R, and within R using the `rjags` package. We will describe both approaches.

Step 1: Data Objects

Specify a 713×20 matrix $\mathbf{X}$ (*without* a leading column of 1's) for individual-level explanatory variables, and a 50×3 matrix $\mathbf{Z}$ for state-level explanatory variables without a leading column of 1's. The contents of these matrices were described in the previous section. Our $\mathbf{Y}$ outcome variable (`grp.influence`) measures the respondents'

perception of interest groups' influence on total budget, special budgets, and general public policies. For modeling purposes it is important to keep these two covariate matrices distinct. All three data structures (the two matrices and the outcome variable vector) are kept in the same JAGS data file where each variable is defined consecutive in the R list format. For example:

```
STATES <- 50
SUBJECTS <- 713
state.id <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
contracting <- c(6, 2, 0, 0, 0, 1, 0, 3, 3, 6, 0, 1, 5, 1, 1, 1, 0, 3, 3, 2,...
gov.influence <- c(19, 13, 11, 21, 17, 19, 15, 14, 19, 17, 13, 15, 6, 19, 13,..
:
nonprofits <-
c(1.9783, 0.509, 2.0701, 1.3639, 15.6682, 2.7968, 2.0128, 0.5585,
7.0653, 3.671, 0.746, 0.7508, 6.6459, 3.61, 2.8795, 1.8049, 1.8861,
1.9391,  0.9202,  3.1633,  3.6798,  4.8333,  3.3067,  1.228,  3.6469,
0.9728, 1.2989, 0.7732, 0.784, 4.2042, 1.0284, 9.8503, 4.2169,
0.5832, 6.4636, 1.9164, 2.1932, 6.6298, 0.7195, 2.1484, 0.6801,
2.9545, 9.9194, 0.8706, 0.5967, 3.9635, 3.5073, 1.1225, 3.4171,
0.4537)
```

This dataset (as well as all of the code) is available at http://dvn.iq.harvard.edu/dvn/dv/jgill, and the code for bringing these data into the R environment for use by rjags is given in Appendix C. Note here that we have embedded the data size constants in the data file. This is much better programming practice than burying these definitions in the actual JAGS code since it makes them more explicit and does not lead to confusion when the code is applied to other data. The variable state.id is used to assign individual cases to the groups they are nested within. This is the essential mapping from the administrator to his/her state of employment. The long series of 1's above indicates that the first set of respondents is all in state 1. Each of the $\mathbf{Z}$ variables has this characteristic: many repeated values indicating sharing of the state assignment, which defines the hierarchy in the data. Notice that the partial listing of the data above finishes with the last group-level variable of length 50.

Step 2: Statistical Specification

We relate these variables through the linear multilevel model:

$$Y_i \sim N(\alpha_{j[i]} + \mathbf{X}_i\beta, \sigma_y^2), \text{ for } i = 1,\ldots,713$$
$$\alpha_j \sim N(\mathbf{Z}\gamma, \sigma_\alpha^2), \text{ for } j = 1,\ldots,50, \quad (9)$$

The first line states that the outcome variable for the ith individual case is distributed normal around a mean-defining systematic component, $\alpha_{j[i]} + \mathbf{X}_i\beta$, with variance σ_y^2. This is exactly the linear model based on the standard Gauss–Markov assumptions except that a distributional feature is assumed, giving the Bayesian context. The matrix $\mathbf{X}$ is defined above and β will be the corresponding estimated regression coefficients. Using hierarchical notation from Gelman and Hill (2007) to indicate that the ith respondent is nested in the jth state, $\alpha_{j[i]}$, we specify a state-specific random

intercept. Therefore there will be $j = 1,\ldots,50$ "intercepts" in the model, each one corresponding to a US state in the data, and denoted at the group level as α_j. The second stage of the Bayesian hierarchical model is to specify a distributional structure for these random effects. This is an assumed distribution from which each of the α_j is drawn. Thus the US states share a common feature, a normal distribution specified in the second line of equation (9), but are represented in intercept terms by distinct draws from this distribution. The power of this model comes from the ability to add variable definitions at this second level in the hierarchy, and here we parameterize at the second level the three explanatory variables in $\mathbf{Z}$ and their corresponding estimated coefficients, γ.

It is very important to note the two different variances accounted for in equation (9). The term σ_y^2 measures the *within-state* variance of the outcome variables (which is assumed to be the same across individuals in different states), whereas the term σ_α^2 gives the variance of the mean estimates *between states*. The first variance term functions as the regular variance of the regression and serves as a reference point. The second term measures differences between states and therefore gives the value of having a multilevel specification defined by state membership. Surprisingly, this is a variance term we would like to be large relative to σ_y^2, since it "soaks up" variability that would normally fall to the error term in the model.

There are two basic blocks of code that users specify in a `JAGS` or `WinBUGS` program: looping through data and parameters applying the link function to relate to the outcome variable, and specifying priors. The first is analogous to specifying a likelihood function, and the second gives distributional assumptions for the unknown parameters. Interestingly, within these blocks the order of individual statements is not important, and the order of the blocks is also not important. This is quite different than programming in R, FORTRAN, C, or some other standard serially specified programming language. In the `JAGS` code we get the specification in equation (9) by first looping through the individual-level variables in $\mathbf{X}$ with:

```
for (i in 1:SUBJECTS) {
  mu[i] <- alpha[state.id[i]]
    + beta[1]*contracting[i] + beta[2]*gov.influence[i] + beta[3]*leg.influence[i]
    + beta[4]*elect.board[i] + beta[5]*years.tenure[i] + beta[6]*education[i]
    + beta[7]*party.ID[i] + beta[8]*category2[i] + beta[9]*category3[i]
    + beta[10]*category4[i] + beta[11]*category5[i] + beta[12]*category6[i]
    + beta[13]*category7[i] + beta[14]*category8[i] + beta[15]*category9[i]
    + beta[16]*category10[i] + beta[17]*category11[i] + beta[18]*category12[i]
    + beta[19]*med.time[i] + beta[20]*medt.contr[i]
  grp.influence[i] ~ dnorm(mu[i],tau.y)
}
```

This loop associates each data value with its corresponding coefficient estimates, looping through each individual case. The random-effects specification, $\alpha_{j[i]}$, is given in the statement `alpha[state.id[i]]`, where the `state.id` variable maps the i individuals to their corresponding state (group), and this is embedded in the alpha vector which is of length 50. Notice that these terms are additively collected in the placeholder `mu[i]` for the ith individual. The last line in the loop specifies that the outcome variable for the ith case is distributed normal around this individual-level

mean with precision (1/variance) `tau.y`. Because this last term is not indexed, it corresponds to all cases in the data, thus giving an estimate of $1/\sigma_y^2$.

Since we want to also give the group-level effects with explanatory variable matrix **Z**, the next part of the code loops through the groups:

```
for (j in 1:STATES) {
    eta[j]    <- gamma[1]*gov.ideology[j] + gamma[2]*lobbyists[j]
              + gamma[3]*nonprofits[j]
    alpha[j] ~ dnorm(eta[j],tau.alpha)
}
```

Here we collect the linear additive components for $\mathbf{Z}\gamma$ in the placeholder `eta[j]` in the first line within the loop. The second line states that the random effect for state j is distributed normal around this state-level mean, `eta[j]`, with precision (1/variance) `tau.alpha`. Because the "connector variable" of `alpha[]`, `JAGS` understands how to relate the individual-level model to the group-level model, achieving the multilevel nesting of individuals into states as specified in equation (9).

Step 3: Prior Definitions

For convenience and model flexibility, we restrict our distributional choice for priors to normal forms. This provides a "conjugate" specification throughout, meaning that the distributional form of the prior flows through to the posterior, albeit with different parameterizations. Conjugacy is a joint property of the prior distribution and the likelihood function such that distributional family of the prior is the same as the posterior with parameters that differ from the information in the likelihood function, where the extent of this change is called "shrinkage."[4] In the linear model with reasonable data size, normal priors produce normal posteriors. This is supported by the standard regression theory for $n = 713$ and the increased numerical stability that is provided by this choice. The set of priors for the individual-level and the group-level coefficients are summarized by:

$$\begin{aligned} \beta &\sim N(\beta_{ky}, \Sigma_\beta) \\ \gamma &\sim N(\gamma_{ky}, \Sigma_\gamma), \end{aligned} \tag{10}$$

where the Σ_β and Σ_γ matrices are diagonal forms with large variances relative to the Kelleher and Yackee's (2009) point estimates. It is possible to specify off-diagonal entries for these matrices, if important correlational information exists before the data analysis, but we saw no reason to add this feature. Such prior information would come from theories about relationships between the coefficients that might exist in the literature under study.

We choose to use informed versions of the prior distributions for some of the unknown parameters since a high-quality source exists for these. Our prior distributions in these cases are diffuse normals centered at the point estimates from Kelleher

4 For instance, a beta distribution prior for an unknown probability parameter plus a likelihood function from a binomial assumption gives a different beta distribution for the posterior of this parameter. The normal distribution is conjugate to itself, as shown in the application here.

and Yackee's (2009, 593) Model 3. We deviate from their values in two important ways: We substitute the variable `elect.board` for their variable `Merit Position`, and since we are using 2008 data only there cannot be a dummy variable for the year 2004 versus 1998. The remaining variables where prior information is low are assigned normal distributions with mean zero and large variance relative to their non-Bayesian model standard errors in previous research. These priors are specified in the `JAGS` code with the sample statements:

```
beta[1] ~ dnorm(0.070,1) # PRIOR MEANS FROM KELLEHER AND YACKEE 2009, MODEL 3
beta[2] ~ dnorm(-0.054,1)
:
beta[9] ~ dnorm(0.0,1) # DIFFUSE PRIORS
beta[10] ~ dnorm(0.0,1)
:
```

(the full set of priors is specified in Appendix C). In some cases we could have looped through distributional assignments for these coefficients that are the same, but since our focus was on thinking carefully about the inclusion of substantive prior information from the literature, we thought it made sense to individualize these decisions on separate lines of code above. In Appendix C we show the complete `JAGS` model specification combining the model statements in the previous paragraphs with the prior statements just above.

ESTIMATION AND RESULTS

We now turn to producing marginal posterior distributions for each of the coefficients of interest using Gibbs sampling as implemented in `JAGS`. The power of using this software is that we do not have to specify each of the full conditional distributions discussed in the previous section. Instead we have the luxury of making software statements, as just described immediately above, that resemble the model statements in equation (9). The joint posterior distribution before commencing this process is a combination of the linear hierarchical likelihood function and the priors, as described in a previous section, given by:

$$\pi(\beta,\gamma,\alpha\,|\,\mathbf{X},\mathbf{Z},\mathbf{Y}) \propto L(\mathbf{X},\mathbf{Z},\mathbf{Y}\,|\,\beta,\gamma,\alpha)p(\beta)p(\gamma)p(\alpha)p(\sigma_y^2)p(\sigma_\alpha^2). \tag{11}$$

This is a 76-dimensional distributional form (20 individual-level coefficients, 3 group-level coefficients, 50 random effects, 2 variances components, and the resulting posterior density), so the analytical process of producing marginal summaries for a regression would be to integrate across 75 dimensions (minus density). Obviously this is impractical, and it is the MCMC process described in the previous section that saves us from attempting this task with calculus tools.

Step 4: Running the Sampler

It is essential to start a Markov chain from some point in the multidimensional space so that it may begin its run. Such a point is theoretically unimportant (Gill 2008, chapter 11) but practically important for one of the standard convergence diagnostics. The

JAGS program when run in terminal model by itself requires that starting values be given in the R vector format, one specification to a line. We ran three parallel chains for each of the 10 imputed datasets where the three chains in each case started from different points. For example, one starting point file contained:

```
tau.y <- 10
tau.alpha <- 10
alpha <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
           1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
beta <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
gamma <- c(1,1,1)
```

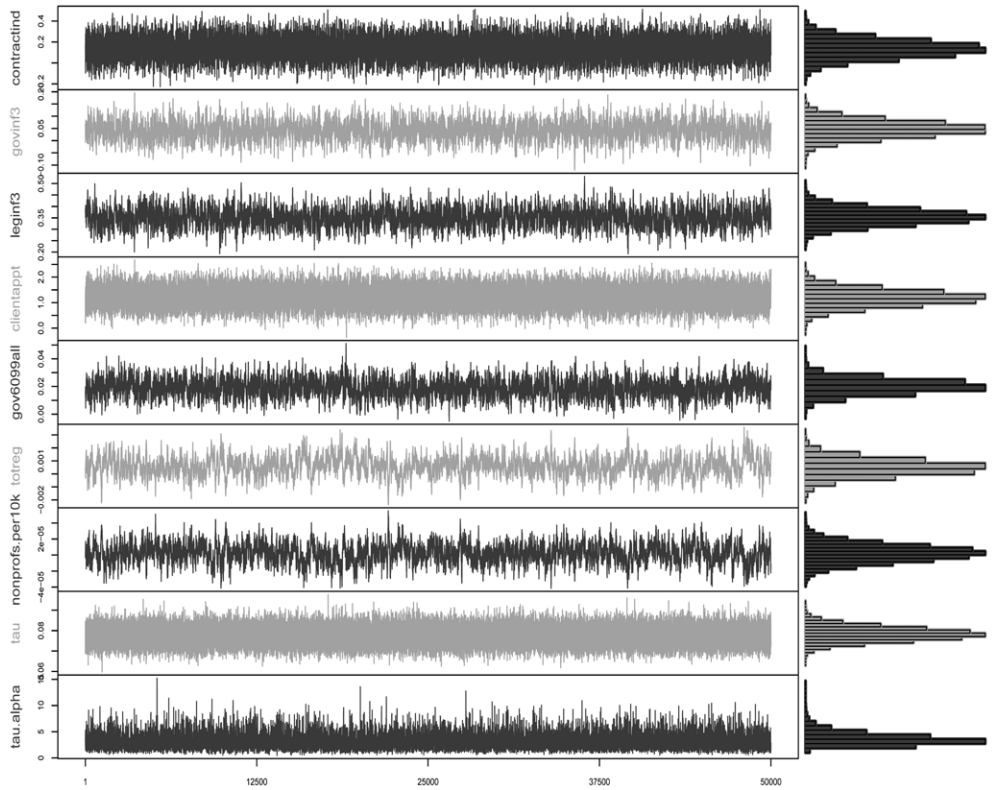
The order of the variables listed does not matter, but each estimated coefficient must have a starting point for MCMC estimation iterations. Appendix C shows how to keep all of these definitions in the R environment rather than specifying files for JAGS run separately.

Using JAGS we ran the Gibbs sampler for 500,000 iterations, keeping only the last half of the run, and therefore using the first half of the run for burn-in. The mechanical details of calling JAGS from the R environment are summarized in Appendix C along with Web site support so that interested readers can reproduce our results immediately. Using the second half of the run, we then ran the diagnostic suite provided by the R package superdiag (Gill and Tsai 2012), which calls all of the conventional convergence diagnostics used in typical MCMC output assessment. This provided no evidence of nonconvergence, giving us confidence that the empirical summary from the path of the last 250,000 iterations converged to the stationary distribution, and thus sufficiently explains the marginal posterior distributions of the parameters of interest. In addition, traceplots that map the traveled path of some critical parameters are given in figure 3 for the last 50,000 iterations along with corresponding histograms. These patterns are typical of all of the marginal results achieved. Notice the stable mean structure in the figure, looking horizontally, for each. In a few cases (gov.ideology, lobbyists, nonprofits), there is a little bit of "snaking," indicating a slower mixing through the sample space than the others. This does not alter our results inferentially; it simply means that we want to make sure that these marginal dimensions are adequately explored by running the Markov chain long enough. All of the formal diagnostics in superdiag point towards convergence and extensive exploration of the posterior space.

Step 5: Summarizing Marginal Posteriors

The coefficient summaries provided in table 1 are the estimated marginal posterior means, standard errors, and 95% credible intervals for the coefficients of the model. These were produced by summarizing the marginal draws from the last 250,000 iterations of the Gibbs sampler, which was done 10 times corresponding to 10 replicate datasets from the multiple imputation process. Bayesian results are often given in probability statements and graphical depictions (see below), but most social science readers still expect to see something that closely resembles a standard regression table. It is important to recall that posterior means, as given here, are not the same estimates as the mode of a likelihood function. However, in applications with only mildly

Figure 3
Traceplots of Sample Draws for Selected Parameters



informative priors and large samples, the difference will be negligible. Furthermore, the same distinction exists between the posterior standard error and the error that results from the square root of the diagonal of the negative inverse Hessian matrix. So table 1 can be read and understood in exactly the same substantive way as a table derived from maximum likelihood estimation, except in cases where the prior is influential and the sample size is small. In this latter case, the assumptions will play an important role and need to be extensively justified.

The JAGS output is a set of marginal posteriors. In R we can simply summarize these empirical draws as if they were data to produce the values in table 1. This is done by:

```
start <- nrow(full.out1)/2; stop <- nrow(full.out1)
asap.out <- cbind(
    apply(full.out1[start:stop,],2,mean),
    apply(full.out1[start:stop,],2,sd),
    apply(full.out1[start:stop,],2,mean)-1.96*apply(full.
    out1[start:stop,],2,sd),
    apply(full.out1[start:stop,],2,mean)+1.96*apply(full.
    out1[start:stop,],2,sd))
```

Table 1
Posterior Summary: Bayesian Hierarchical Model, State Lobbying

Parameters	Mean	Std. Error	95% Lower CI	95% Upper CI
α mean(1:50)	1.3905	0.7037	0.0112	2.7698
contracting	0.1987	0.0963	0.0099	0.3874
gov.influence	0.0481	0.0367	−0.0239	0.1202
leg.influence	0.3519	0.0397	0.2741	0.4297
elect.board	1.3436	0.3546	0.6486	2.0386
years.tenure	0.0347	0.0233	−0.0110	0.0804
education	0.1249	0.1217	−0.1136	0.3634
party.ID	−0.0046	0.0845	−0.1703	0.1611
category2	−0.4282	0.5423	−1.4912	0.6348
category3	−0.0596	0.5885	−1.2131	1.0938
category4	1.5501	0.4571	0.6541	2.4461
category5	−0.5473	0.5010	−1.5292	0.4347
category6	0.9227	0.5395	−0.1348	1.9801
category7	1.7014	0.4353	0.8482	2.5546
category8	1.0013	0.4986	0.0240	1.9785
category9	0.9412	0.4860	−0.0115	1.8938
category10	0.6157	0.4634	−0.2925	1.5239
category11	−0.1264	0.4265	−0.9624	0.7096
category12	−0.1592	0.5727	−1.2816	0.9632
med.time	1.1435	0.3587	0.4405	1.8465
medt.contr	−0.0869	0.1372	−0.3559	0.1821
gov.ideology	0.0182	0.0062	0.0060	0.0303
lobbyists	0.0007	0.0008	−0.0007	0.0022
nonprofits	−0.0217	0.1267	−0.2701	0.2266
τ_y	0.0763	0.0042	0.0682	0.0845
τ_α	3.1021	1.3523	0.4517	5.7525

where `full.out1` is one of our 10 replicated results due to the multiple imputation process. This setup assumes that `JAGS` was run in terminal mode and the samples loaded into R with `read.coda`. Notice that all we are doing here is calculating and manipulating means and standard deviations for last half of the rows of the data matrix that contains the MCMC simulations. Some prefer to obtain such tabular results from the MCMC analysis suites in R: `coda` and `boa`, which are menu-driven and quite simple to use. The key point here is that once the Gibbs sampler is done and we have confidence that it has converged and fully explored the sample space, the rest of the analysis is as easy as a standard data summary. This is one very nice feature of working with MCMC results.

One final note about reporting results is important here. Notice from table 1 that we did not report the marginal posterior summary for 50 α coefficient values corresponding to the 50 estimated random effects for the US states. Obviously this would have made the table much larger and perhaps more cumbersome. Many authors, therefore, choose to summarize them graphically or provide and discuss important cases from amongst the groups. In our case we only showed the Markov chain path for α_1 and α_2 in figure 1, although it would have been straightforward to present more information about all of the estimated random effects. The mean of these effects is

1.3905, which shows a positive effect on the outcome variable even for zero levels at all of the explanatory variables and at the reference level ("other") of agency category.

Step 6: Assessing Model Fit

Model fit statistics flow naturally from the MCMC output. First we can compare deviance from the fit model to the null (intercept only) model in exactly the same fashion as regular generalized linear model analysis. The summed deviance function comes from comparing the predicted value of the outcome value to actual value, which is just a residuals analysis with linear models. We get this easily from the `JAGS` or `WinBUGS` output by setting a monitor to accumulate the deviance at each step of the Markov chain (`monitor set deviance` in the `JAGS` terminal window) and then simply taking the mean across iterations. As a reminder, the model deviance is compared to the saturated model deviance of zero and the null model deviance, where differences are asymptotically distributed χ^2 with degrees of freedom equal to the difference in the number of parameters. So for our model, we see the results in table 2 where the χ^2 difference between models produces right-hand-side tail values that are essentially zero, indicating evidence that the estimation differs from both the null model and the saturated model. This indicates that our model is statistically distinct from both the null model and the saturated model, meaning that we have substantial progress away from the simplest modeling approach but we still have unexplained variance relative to a fully saturated specification.

The standard errors from the τ values $(\sqrt{1/\tau})$ in table 1 are $\sigma_y = 3.6202$ and $\sigma_\alpha = 0.5677$, which defends the use of a hierarchical model since there is an appreciable amount of model variability soaked up in the between-state specification that would have otherwise defaulted to regression standard error in a nonhierarchical specification. Normally we would like to see a larger value of σ_α relative to σ_y, perhaps even substantially greater, but being a similar order of magnitude easily justifies the multilevel definition (see the further practical guidance in Gelman and Hill 2007).

We also want to compare our model to the straw-man specification of the null model in terms of the DIC given in equation (8). Recall that this is a Bayesian analog to the AIC that accounts for the varying roles that parameters play in a hierarchical model. The DIC value for our model is obtained by rerunning the simulations recording the $\bar{D}$ and p_D values, according to the following command in R (terminal-run `JAGS` just inserts an additional model command),

```
asap.dic <- dic.samples(asap2.model, n.iter=250000, type="pD")
```

which reports $DIC = \bar{D} + p_D = 3861 + 34.04 \approx 3895$. This process is repeated for the null model (one devoid of explanatory variables other than the random effect α, but

Table 2
Deviance Analysis, χ^2 Statistics

Model	Deviance	Difference	DF	Tail Value
Null	3963.8	103.7	24	6.9787e-12
Estimated	3861.1			
Saturated	0.0	3861.1	689	<1.0e-300

using the exact same data). The same `dic samples` command on this smaller model produces $\bar{D} = 3964$ and $p_D = 49.26$. The null model DIC is the rounded up sum of these two values: 4014. From this we obviously prefer the fully specified model since it has lower DIC (3895 versus 4014). In more general circumstances, we might be comparing different mixes of covariates in the specification, and the DIC would be a useful guide to relative model quality between these specifications. In this example our strong theoretical focus on extending the Kelleher and Yackee (2009) model led us to a single well-defined model specification, with only a DIC comparison to the null concerning us.

It is also possible to produce very interesting and useful visualizations of the marginal posterior distributions and their implications. For instance, we can generate *posterior predictive distributions* of the data from the fitted model and compare these to the observed data (Gelman and Hill 2007, chapter 24; Gill 2008, chapter 6). Graphing the marginal posteriors and showing interesting features, including cut-points and dispersion, can also reveal important substantive conclusions that would complement the results in table 1.

Discussion of Findings

Our findings using the 2008 data and the priors from the earlier analyses provide largely consistent results regarding interest group influence in state agencies. First, for the core concern of Kelleher and Yackee's (2009) article, we also observe that more contracting is positively related to perceptions of interest group influence, and that this finding is statistically reliable, having a credible interval of [0.0099 : 0.3874]. We also find clear evidence that those agency heads who spent above the median amount of time with organized interests perceived groups to have more influence over their agencies (a posterior mean of 1.1435), which differs from Kelleher and Yackee's (2009) results but is consistent with their earlier work where they found that time spent with interest groups was an important predictor of their perceived influence using a continuous measure of time (Kelleher and Yackee 2006). In contrast with Kelleher and Yackee's (2009) findings, we do not find strong evidence that the interaction between time and contracting is a reliable predictor of perceived interest group influence. The posterior mean point estimate for this variable is actually negative (–0.0869), and the posterior variance is sufficiently large (0.1372) that we do not consider this to be a reliable find: 41% of the posterior density is below zero and 59% of the posterior density is above zero, under a normal assumption (an assumption justified by both the data size and the MCMC estimation process).

Several of the other results are also quite interesting from a substantive standpoint. First, consistent with Brudney and Hebert (1987), we find that agency heads who were elected or appointed by boards or commissions perceive interest group influence to be much greater, and this coefficient is reliable: The posterior distribution 95% credible interval is [0.6486 : 2.0386]. This likely indicates that the more important the role that interest groups play in these selection processes, the more influence over administrators occurs once they are selected. One of the more interesting findings is that a legislature perceived to be more influential is associated with *more*

powerful clientele groups: The posterior distribution 95% credible interval in table 1 is [0.2741 : 0.4297], which mirrors Kelleher and Yackee's (2009) findings. This almost certainly reflects the idea that legislators intervene in the bureaucracy on behalf of organized interests, or administrators consider the preferences of legislatures' interest group allies when making decisions, leading to complimentary interest group and legislative influence (Witko 2011). We also observe that more time in the current position is associated with the perception of greater interest group influence, but this coefficient is not as reliable as we would like: The posterior mean of 0.0347 and posterior standard deviation of 0.0233 imply that 7% of the posterior density is below zero under a normal assumption. This is a classic example where the Bayesian approach provides a more flexible reporting mechanism. In standard statistical analysis with the typical threshold of $\alpha = 0.05$ (two-tailed as is the custom with regression results), we would be forced to dismiss this effect. With the Bayesian way of thinking, there is a 93% chance (.93 probability) that job tenure is positively associated with greater interest group influence. Notice that it is now entirely left to the reader to determine whether this is sufficient evidence to convince them, rather than force adherence to a completely arbitrary threshold. We also find evidence (a posterior credible interval of [0.0060 : 0.0303]) that in states with more liberal governments agency heads tend to perceive interest groups as having more power. It is not obvious why this is the case, and this finding suggests a direction for future research.

Kelleher and Yackee's (2009) argument regarding contracting and interest group influence was intriguing but had received little empirical attention. Using a new round of data and incorporating the information from their study, we find that more contracting is associated with more perceived interest group influence over agency policies and budgets. Based on our analysis we can also make some broader statements about the influence of interest groups in administrative processes. Because the contracting relationships and spending more than the typical amount of time with interest group leaders is associated with more perceived influence, we can conclude that direct access enables interest group influence over the decision making of public managers, probably in much the same manner that legislators are influenced by interest groups. Thus, our findings further highlight some of the potential downsides of decentralized governance networks that involve close collaborative relationships between public managers and representatives of external groups (Kelleher and Yackee 2009; O'Toole and Meier 2006; Whitford 2002; Witko 2011). However, our finding that more legislative power is also associated with greater perceived interest group influence indicates that, whatever the other benefits of increased centralization (Whitford 2002), more centralization would perhaps not limit interest group influence over agencies because groups can mobilize central political principals to pressurize agencies on their behalf (Stigler 1971). A major task of future research is to determine how different levels of decentralization and centralization may condition group influence in the bureaucracy.

One major benefit of the Bayesian approach to explore these types of questions going forward is that there is a clear process for explicitly incorporating the information from other studies. It does not make sense to conclude that previous findings are somehow "wrong" based on one additional study. Instead we should use the existing evidence and our new data to update our understanding of a phenomenon. After doing this here, there does appear to be a relationship between contracting and group

influence. Future studies of federal, local, or state governments using different data sources should directly incorporate the findings of these previous studies into their analyses of new data.

29.5 CONCLUSION

The Bayesian process of data analysis 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 overt in the Bayesian setup, readers can more accurately assess model quality.

Bayesian statistical models cannot be developed in a "cookbook" fashion, looking up the recipe steps and blindly following the instructions, such as the statistical testing recipes that are on the inside cover of some basic texts. Instead, each *step* of the process, from determining the form of prior distributions and likelihood function through estimation and description of results, needs to be done in a thoughtful and deliberate way so that the decisions are clear to readers and the results have integrity. We provided the "steps" here not as confining mechanism but as a convenient reminder of the process. Bayesians carefully follow this process for historical reasons since there was about 100 years of hostility from others, leading to a defensiveness whose manifestation was the detailed justification of every model assumption. It turns out that this is a prescription for all developers of statistical models, and the routinization of model building harms many disciplines including public administration. We feel that injecting a Bayesian inferential culture into empirical research in public administration moves us to a better scientific process of discovery and away from lockstep procedures followed by convention, as routinely done in closely aligned disciplines such as the study of business administration.

Social scientists have increasingly embraced Bayesian methods as useful ways to address empirical and methodological problems. Over the last two decades, any sense of controversy has receded from the field of statistics, with plenty of evidence in top statistics journals. Now with a wide range of freely available MCMC tools, estimation challenges are fairly easily managed, even under seemingly difficult circumstances. This leads to a world where public administration scholars have few impediments to developing useful and principled Bayesian models for their empirical questions. We demonstrated the utility of this approach here, with an important substantive application that sheds light on a broad question in public administration.

We have no illusion that public administration scholars are going to transform into research statisticians. Nor do we believe that they should. Our colleagues justifiably care about theories of government and administration, providing evidence to support these through collected data and placing their conclusions into our rich literature. The statistical modeling part of this process should not be viewed as an annoying

encumbrance, but should instead be considered an opportunity for creative exploration. The discipline subsumes this fertile inquiry by locking the empirical process into old practices that slow growth in the field.

Bayesian methods are not a panacea for all quantitative work. It is still possible to construct flawed Bayesian specifications just as it is possible to construct flawed non-Bayesian specifications. The researcher still needs to carefully consider the structure of the statistical model with regard to data measurement, parameter relationships, and descriptions of uncertainty. Still, the Bayesian paradigm provides a more principled approach to describing uncertainty from data and models. As Ed George observed, "All good procedures are Bayesian, but not all Bayesian procedures are good" (personal communication). We believe that the science of public administration can be improved by the more appropriate view of probability and uncertainty contained in the Bayesian paradigm.

APPENDIX A. Data Description

This section summarizes the data format and coding decisions applied to the final subset of the 2008 ASAP data. Missing data (2.75% here from our original version of the 2008 data with 47 variables) were imputed with the mice package in R that uses multiple imputation. Ten replicate datasets were created, each of which is fully complete after the process. Each of the 10 replicate datasets were subsetted from 47 original variables down to the 13 explanatory variables and 1 outcome variable used in the model. As described above, this became a 713×20 indvidual-lavel explanatory variable matrix $\mathbf{X}$ due to the treatment contrast for category_[], and a 50×3 group-level explanatory matrix $\mathbf{Z}$. We then ran the Gibbs sampler on each replicate and averaged the posterior means to create a regression table. The posterior standard errors are a weighted average of between- and within-replication uncertainty (Rubin 1987). The subsetted dataset (14 variables) and a larger version of the dataset (47 variables) are available for replication purposes at the authors' dataverse page: http://dvn.iq.harvard.edu/dvn/dv/jgill. Several of the 47 variables are embargoed for confidentiality purposes at the request of the original collector of the data, although not any of the variables used here.

1. grp.influence is a scale from 3:21 created from adding three seven-point scales: respondents' perceptions of the influence that clientele groups have on the total agency budget, specialized program budgets, and agency policies. The observed distribution of these outcomes is 3(38), 4(19), 5(33), 6(60), 7(37), 8(66), 9(57), 10(72), 11(64), 12(72), 13(53), 14(39), 15(43), 16(23), 17(13), 18(14), 19(7), 20(2), 21(1) denoting *code(count)*.

2. contracting is a scale from 0:6 with observed distribution [263, 142, 112, 60, 37, 27, 72]. Higher levels on this scale indicate more private contracting within the respondent's agency.

3. gov.influence is the respondents' assessment of the governor's influence on contracting in his/her agency. This variable ranges from 0:21 with observed distribution [4, 1, 2, 14, 19, 13, 18, 12, 18, 29, 40, 31, 41, 72, 65, 102, 77, 44, 111].

4. leg.influence is the respondents' assessment of the legislatures' influence on contracting in his/her agency, ranging from 0:21 with observed distribution [2, 1, 3, 4, 4, 15, 10, 16, 25, 22, 44, 51, 64, 60, 82, 112, 70, 47, 81].
5. elect.board is a dichotomous variable coded 1 if appointed by a board, a commission or elected, and 0 otherwise. The distribution seen with this sample is [577, 129].
6. years.tenure gives the number of years that the respondent has worked at their current agency. The observed mean is 5.6752 with standard deviation 5.6845.
7. education is a ordinal variable for level of education possessed by the respondent: high school (6 cases), some college (27 cases), bachelors degree (139 cases), graduate study (80 cases), and graduate degree (449 cases).
8. partisan.ID is a five-point ordinal variable (1–5) for the respondent's partisanship. It is scaled: strong Democrat (310), weak Democrat (84), Independent (60), weak Republican (55), and strong Republican (172). The 10 cases coded as "don't know" or "refuse" were given missing value indicators and imputed.
9. category_[] categorizes the agency type. The observed distribution of the 13 types is [38, 42, 30, 74, 52, 43, 80, 52, 56, 66, 87, 33, 55]. With "other" being the reference type, we scramble the order of these in the data to remove substantive labels for privacy reasons.
10. med.time indicates whether the respondent spent more or less than the sample median with representatives of interest groups. There were 442 less-than or equal-to-the-median cases (coded 0) and 271 more-than cases (coded 1).
11. medt.contr is a created interaction variable crossing med.time with contracting.
12. gov.ideology is the state government ideology from Berry et al. (1998), which ranges from 0 to 100. When used in our analysis we observed in this sample a mean of 48.018 and a standard deviation of 25.976. To access updated data: http://www.bama.ua.edu/rcfording/stateideology.html.
13. lobbyists is the total state lobbying registrants in 2000–1 from Gray and Lowery (1996, 2001). Here we observed a mean of 670.84 and a standard deviation of 425.89.
14. nonprofits provides the total number of nonprofit groups in the respondents' state in the year 2008. Since the number is very large we divided it by 10,000 to put it on a meaningful scale in the model results. The sample has a mean of mean of 2.6833 and a standard deviation of 2.5259 (scaled).

APPENDIX B. MCMC Estimation Theoretical Background

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 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., like desired posteriors), we start a Markov chain in the appropriate sample space and allow it to 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 Gibbs sampler variant of MCMC (Geman and Geman 1984) works as follows. For convenience define $\phi = [\beta, \gamma, \mathbf{b}, \tau]$ as the vector of unknown parameters. Call $\phi_{[i]}$ the ϕ vector where the ith parameter is removed 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. 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 to a limiting distribution that is the target posterior, provided that the chain is—ergodic A chain is ergodic if all of its states are ergodic—aperiodic and positive recurrent. A sufficient condition for aperiodicity is that the probability of remaining in the same state is nonzero (discrete chains) or the probability of remaining in the same region is nonzero (continuous chains): $P(\mathfrak{X}, \mathfrak{X}) > 0$. A state is positive recurrent if the mean time to transition back to the same state is bounded. A chain is positive recurrent if and only if it has a stationary distribution, $\pi, \mathfrak{Z}: n \overset{\lim}{\to} \infty \sum_{k=1}^{n} P^k(\mathfrak{Y}, \mathfrak{Z}) = \pi(\mathfrak{Y})$ for all Y and Z in the parameter space. 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 converges almost surely to strongly consistent estimates of the parameters of the limiting distribution, despite dependence on some state space $S \in \Re$ for a given transition kernel and initial distribution. These properties for the Gibbs sampler are well studied and are not provided here. See Carlin and Louis (2000), Gamerman and Lopes (2006), Gelfand and Smith (1990), Gelman and Rubin (1992), Gelman et al. (2003), Geweke (1989), Tanner (1996) (to name just a few) for excellent comprehensive discussions of the theory and practice of Gibbs sampling and MCMC in general. The original article on the statistical application of Gibbs sampling (Geman and Geman 1984), however, is a far more demanding read and applies the algorithm to photo image restoration.

An added wrinkle is required in this application because the complete conditionals for γ parameters do not have an easily obtainable form. In these cases an update is produced for these parameters using a modified Metropolis–Hastings (Hastings 1970; Metropolis and Ulam 1949; Metropolis et al. 1953) step within the Gibbs sampling (Cohen et al. 1998; Gamerman and Lopes 2006). This works as follows. First transform γ_j ($j \in J$ parameters in γ such that it has support over $\Re$) and draw a normal approximation, $\hat{\gamma}_j$, centered at the current value of γ_j in the simulation and using the variance from past iterations, $s^2_{\gamma_j}$ (using 1 as a starting value). Then at the kth iteration, take a draw from the conditional $\pi(\hat{\gamma}_i \mid \phi_{[j]})$ and make the following transition to the $(k+1)^{\text{st}}$ value for γ_i:

$$\gamma_j^{[k+1]} = \begin{cases} \hat{\gamma}_j \text{ with probability } P\left(\min\left(\dfrac{\pi(\hat{\gamma}_j \mid \phi_{[j]})}{\pi(\gamma_j^{[k]} \mid \phi_{[j]})}, 1 \right) \right) \\ \gamma_j^{[k]} \text{ with probability } 1 - P\left(\min\left(\dfrac{\pi(\gamma_j \mid \phi_{[j]})}{\pi(\gamma_j^{[k]} \mid \phi_{[j]})}, 1 \right) \right). \end{cases} \tag{12}$$

So unlike the Gibbs sampler, the Metropolis–Hastings algorithm does not necessitate movement on every iteration. In fact, it can be shown both that the Gibbs sampler is a generalization of Metropolis–Hastings where the probability of accepting the candidate value is always 1 (Tanner 1996, 182), and that Metropolis–Hastings is a generalization of the Gibbs sampler where movement is not necessary, the full conditionals are not required, and the previous value of the component to be (potentially) updated is consulted (Besag et al. 1995; Gamerman and Lopes 2007).

The "Metropolis within Gibbs" algorithm described here is distinct from two other hybrid techniques and should not be confused with those. One of those approaches is to use the Metropolis–Hastings algorithm only when computationally difficult steps for a parameter are encountered during the run (Gelman 1992). In addition, some authors (Tierney 1991) recommend switching back and forth between the two methods as a way of avoiding areas of the posterior density that are dominated by local maxima. The method applied here uses the hybrid described not for such computational reasons (although they are not precluded) but because the form of the hierarchical model produces a posterior where the full set of conditionals are simply unavailable.

The ergodic theorem shows that after a sufficiently large number of chain iterations are performed, then subsequent draws are from the target limiting posterior distribution: $P(p_i \mid y_i, n_i, \delta_1, \delta_2)$. Reality is rarely this clear. Two primary philosophies compete for adherents among applied researchers. Gelman and Rubin (1992) suggest using the EM algorithm (or some variant) to find the mode or modes of the posterior, then create an over-dispersed 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 (1992) recommends implementing one long chain and using well-known time series statistics used 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 allow one chain to 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.

Another very useful tool is Geweke's (1992) convergence statistic. The idea behind this test is to compare some proportion of the early part of the chain after the burn-in period with some nonoverlapping 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.

APPENDIX C. R Code for MCMC Estimation With JAGS

This appendix provides the R code to set up data structures and run the model with rjags. This process is given for only 1 of the 10 replicate datasets from the multiple imputation process. Our results were almost identical across the 10 sets of marginal posteriors (table 1), but in general the 10 sets of coefficient estimates are averaged and the 10 sets of standard errors are combined by the square root of the sum of the within-model variance plus 10/9 times the between-coefficient variance (Little and Rubin 1983, 2002; Rubin 1987). All of the code below and one of our replicate datasets are available online at: http://dvn.iq.harvard.edu/dvn/dv/jgill.

```
# LOAD REQUIRED LIBRARIES
library(rjags); library(arm); library(coda); library(superdiag)

# LOAD DATA WITH INDIVIDUAL-LEVEL AND GROUP-LEVEL VARIABLES
asap.individual.data <- read.table("Article.JPART/asap.individual.dat",header=TRUE)
asap.group.data <- read.table("Article.JPART/asap.group.dat",header=TRUE)

# CONDITION THE TERMINAL JAGS DATA INTO A LIST FOR rjags
system("echo 'asap.jags.list <- list(' > Article.JPART/asap.rjags.dat")
system("tail -c +31 Article.JPART/asap.jags.2.dat >> Article.JPART/asap.rjags.dat")
system("cat Article.JPART/asap.rjags.dat | sed -e 's/)/),/' > Article.JPART/temp")
system("mv Article.JPART/temp Article.JPART/asap.rjags.dat")
system("echo 'STATES <- 50, SUBJECTS <- 713)' >> Article.JPART/asap.rjags.dat")
source("Article.JPART/asap.rjags.dat")
names(asap.jags.list) <- c("state.id", "contracting", "gov.influence",
  "leg.influence", "elect.board", "years.tenure", "education", "party.ID",
  "category2", "category3", "category4", "category5", "category6", "category7",
  "category8", "category9", "category10", "category11", "category12", "med.time",
  "medt.contr", "grp.influence", "gov.ideology", "lobbyists", "nonprofits",
  "STATES", "SUBJECTS")

# DEFINE THE MODEL
asap.model2.rjags <- function() {
  for (i in 1:SUBJECTS) {
    mu[i] <- alpha[state.id[i]]
  + beta[1]*contracting[i] + beta[2]*gov.influence[i] + beta[3]*leg.influence[i]
  + beta[4]*elect.board[i] + beta[5]*years.tenure[i] + beta[6]*education[i]
  + beta[7]*party.ID[i] + beta[8]*category2[i] + beta[9]*category3[i]
  + beta[10]*category4[i] + beta[11]*category5[i] + beta[12]*category6[i]
  + beta[13]*category7[i] + beta[14]*category8[i] + beta[15]*category9[i]
  + beta[16]*category10[i] + beta[17]*category11[i] + beta[18]*category12[i]
  + beta[19]*med.time[i] + beta[20]*medt.contr[i]
grp.influence[i] ~ dnorm(mu[i],tau.y)
}
for (j in 1:STATES) {
eta[j] <- gamma[1]*gov.ideology[j] + gamma[2]*lobbyists[j]
+ gamma[3]*nonprofits[j]
```

```
alpha[j] ~ dnorm(eta[j],tau.alpha)
}
beta[1] ~ dnorm(0.070,1); # PRIOR MEANS FROM KELLEHER AND YACKEE 2009, MODEL 3
beta[2] ~ dnorm(-0.054,1);
beta[3] ~ dnorm(0.139,1);
beta[4] ~ dnorm(0.051,1);
beta[5] ~ dnorm(0.017,1);
beta[6] ~ dnorm(0.056,1);
beta[7] ~ dnorm(0.039,1);
beta[8] ~ dnorm(0.0,1); # DIFFUSE PRIORS
beta[9] ~ dnorm(0.0,1);
beta[10] ~ dnorm(0.0,1);
beta[11] ~ dnorm(0.0,1);
beta[12] ~ dnorm(0.0,1);
beta[13] ~ dnorm(0.0,1);
beta[14] ~ dnorm(0.0,1);
beta[15] ~ dnorm(0.0,1);
beta[16] ~ dnorm(0.0,1);
beta[17] ~ dnorm(0.0,1);
beta[18] ~ dnorm(0.0,1);
beta[19] ~ dnorm(0.184,1); # PRIOR MEANS FROM KELLEHER AND YACKEE 2009, MODEL 3
beta[20] ~ dnorm(0.156,1);
gamma[1] ~ dnorm(0.0,1); # DIFFUSE PRIORS
gamma[2] ~ dnorm(0.0,1);
gamma[3] ~ dnorm(0.0,1);
tau.y ~ dgamma(1.0,1);
tau.alpha ~ dgamma(1.0,1);
}

# WRITE THE MODEL STATEMENT OUT TO A FILE
write.model(asap.model2.rjags, "Article.JPART/asap.model2.rjags")

# SET THE INITIAL (STARTING VALUES)
inits <- list(tau <- 10, tau.alpha <- 10, alpha <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
beta <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
gamma <- c(1,1,1))

# RUN THE MODEL
asap2.model <- jags.model(file="Article.JPART/asap.model2.rjags",
data=asap.jags.list, n.chains=3, n.adapt=0)
update(asap2.model, n.iter=250000)
asap2.mcmc <- coda.samples(model=asap2.model, variable.names=names(asap.jags.list),
n.iter=250000)
summary(asap2.mcmc)

# CHECK CONVERGENCE
superdiag(asap2.mcmc)

# GET THE DEVIANCE AND THE DIC
    asap2.dic <- dic.samples(asap2.model, n.iter=250000, type="pD")
```

REFERENCES

Akaike, H. 1976. Canonical correlation analysis of time series and the use of an information criterion. In *System identification: Advances and case studies*, ed. R. K. Mehra and D. G. Lainiotis, 52–107. New York: Academic Press.

Berry, William D., Evan J. Ringquist, Richard C. Fording, and Russel L. Hanson. 1998. Measuring citizen and government ideology in the American states, 1960–93. *American Journal of Political Science* 42(1):327–48.

Besag, J., P. J. Green, D. M. Higdon, and K. L. Mengersen. 1995. Bayesian computation and stochastic systems (with discussion). *Statistical Science* 10:3–66.

Birnbaum, A. 1962. On the foundations of statistical inference. *Journal of the American Statistical Association* 57:269–306.

Bowling, Cynthia J., Jennifer Jones, and Christine A. Kelleher. 2006. Cracked ceiling, firmer floors, and weakening walls: Trends and patterns in gender representation among executives in American state governments, 1970–2000. *Public Administration Review* 66(6):823–36.

Bowling, Cynthia J., and Deil S. Wright. 1998. Change and continuity in state administration: Administrative leadership across four decades. *Public Administration Review* 58(5):429–44.

Boyne, George A., Kenneth J. Meier, Laurence J. O'Toole, and Richard M. Walker. 2005. Where next? Research directions on performance in public organizations. *Journal of Public Administration Research and Theory* 15:633–9.

Brudney, Jeffrey L., and Ted F. Hebert. 1987. State agencies and their environments: Examining the influence of important external actors. *The Journal of Politics* 49(1):186–206.

Brudney, Jeffrey L., Sergio Fernandez, Jay Eungha Ryu, and Deil S. Wright. 2005. Exploring and explaining contracting out patterns in the American states. *Journal of Public Administration Research and Theory* 15(3):393–419.

Carlin, B. P., and T. A. Louis. 2000. *Bayes and empirical Bayes methods for data analysis*, 2nd ed. New York: Chapman & Hall.

Cohen, J., D. Nagin, G. Wallstrom, and L. Wasserman. 1998. Hierarchical Bayesian analysis of arrest rates. *Journal of the American Statistical Association* 93:1260–70.

Cranmer, Skyler, and Jeff Gill. 2013. We have to be discrete about this: A non-parametric imputation technique for missing categorical data. *British Journal of Political Science*. Forthcoming.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Statistical Society of London A* 222:309–60.

———. 1925a. *Statistical methods for research workers*. Edinburgh, UK: Oliver and Boyd.

———. 1925b. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* 22:700–25.

———. 1930. Inverse probability. *Proceedings of the Cambridge Philosophical Society* 26:528–35.

———. 1934. *The design of experiments*, 1st ed. Edinburgh, UK: Oliver and Boyd.

Gamerman, D., and H. F. Lopes. 2006. *Markov chain Monte Carlo*, 2nd ed. New York: Chapman & Hall.

Gelfand, A. E., and A. F. M. Smith. 1990. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85:398–409.

Gelman, A. 1992. Iterative and non-iterative simulation algorithms. *Computing Science and Statistics* 24:433–8.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. *Bayesian data analysis*, 2nd ed. New York: Chapman & Hall.

Gelman, A., and J. Hill. 2007. *Data analysis using regression and multilevellhierarchical models*. Cambridge, UK: Cambridge University Press.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* 7:457–511.

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721–41.

Geyer, C. J. 1992. Practical Markov chain Monte Carlo. *Statistical Science* 7:473–511.

Geweke, J. 1989. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57:1317–39.
———. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian statistics 4*, ed. J. M. Bernardo, A. F. M. Smith, A. P. Dawid, and J. O. Berger, 169–93. Oxford, UK: Oxford University Press.
Gill, J. 1999. The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52:647–74.
———. 2008. *Bayesian methods for the social and behavioral sciences*, 2nd ed. New York: Chapman & Hall.
———. 2008. Is partial-dimension convergence a problem for inferences from MCMC algorithms? *Political Analysis* 16:153–78.
Gill, Jeff, and George Casella. 2009. Nonparametric priors for ordinal Bayesian social science models: Specification and estimation. *Journal of the American Statistical Association* 104:453–64.
Gill, Jeff, and John Freeman. 2013. Dynamic elicited priors for updating covert networks. *Network Sciences*. Forthcoming.
Gill, Jeff, and Kenneth J. Meier. 2000. Public administration research and practice: A methodological manifesto. *Journal of Public Administration Research and Theory* 10(1):157–200.
Gill, Jeff, and Lee Walker. 2005. Elicited priors for Bayesian model specifications in political science research. *Journal of Politics* 67:841–72.
Gray, Virginia, and David Lowery. 1996. *The population ecology of interest representation: Lobbying communities in the American states*. Ann Arbor: University of Michigan Press.
———. 2001. The institutionalization of state communities of organized interests. *Political Research Quarterly* 54:265–84.
Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.
Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90:773–95.
Kelleher, Christine A., and Susan Webb Yackee. 2006. Who's whispering in your ear? The influence of third parties over state agency decisions. *Political Research Quarterly* 59(4):629–43.
———. 2009. A political consequence of contracting: Organized interests and state agency decision-making. *The Journal of Public Administration Research and Theory* 19(3):579–602.
Kettle, Donald F. 2000. The transformation of governance: Globalization, devolution and the role of government. *Public Administration Review* 60:488–97.
Little, R. J. A., and D. B. Rubin. 1983. On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *The American Statistician* 37:218–20.
———. 2002. *Statistical analysis with missing data*, 2nd ed. New York: John Wiley & Sons.
Mackenzie, Calvin G., and Paul Light. 1987. *Presidential Appointees, 1964–1984 (ICPSR Study 8458)*. Ann Arbor, MI: Intra-University Consortium for Political and Social Research.
Meier, Kenneth J., and Laurence J. O'Toole, Jr. 2006. Political control versus bureaucratic values: Reframing the debate. *Public Administration Review* 66(2): 177–92.
Metropolis, N., and S. Ulam. 1949. The Monte Carlo method. *Journal of the American Statistical Association* 44:335–41.
Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* 21:1087–91.
Miller, Cheryl M. 1987. State administrator perceptions of the policy influence of other actors: Is less better? *Public Administration Review* 47(3):239–45.
Miller, Cheryl M., and Deil S. Wright. 2009. Who's minding which store? Institutional and other actors' influence on administrative rulemaking in state agencies, 1978–2004. *Public Administration Quarterly* 33(3):397–428.
Moe, Terry M. 1985. Control and feedback in economic regulation: The case of the NLRB. *The American Political Science Review* 79(4):1094–116.
Neyman, J., and E. S. Pearson. 1928a. On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika* 20A:175–240.

———. 1928b. On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika* 20A:263–294.

———. 1933a. On the problem of the most efficient test of statistical hypotheses. *Philosophical Transactions of the Royal Statistical Society, Series A* 231:289–337.

———. 1933b. The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society* 24:492–510.

———. 1936. Sufficient statistics and uniformly most powerful tests of statistical hypotheses. *Statistical Research Memorandum* 1:113–137.

O'Toole, Laurence J., Jr., and Kenneth J. Meier. 2006. Networking in the penumbra: Public management cooptative links and distributional consequences. *International Public Management Journal* 9(3):271–94.

Potoski, Matthew. 1999. Managing uncertainty through bureaucratic design: Administrative procedures and state air pollution control agencies. *Journal of Public Administration Research and Theory* 9(4):623–40.

Raftery, A. E. 1995. Bayesian model selection in social research. *Sociological Methodology* 25:111–64.

Reenock, Christopher M. and Brian J. Gerber. 2008. Political insulation, information exchange, and interest group access to the bureaucracy. *Journal of Public Administration Research and Theory* 18:415–440.

Rubin, D. B. 1987. *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

———. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 91:473–89.

Samaniego, Fancisco J. 2010. *A comparison of the Bayesian and Frequentist approaches to estimation*. New York: Springer-Verlag.

Schafer, J. L. 1997. *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Spiegelhalter, D., Best, N. G., Carlin, B., and van der Linde, A. 2002. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society, Series B*, 64: 583–640.

Stigler, George. 1971. The theory of economic regulation. *Bell Journal of Economics and Management Science* 2:3–21.

Tanner, Martin. 1996. *Tools for statistical inference*, 3rd ed. New York: Springer-Verlag.

Tierney, L. 1991. Exploring posterior distributions using Markov chains. In *Computing science and statistics: Proceedings of the 23rd symposium on the interface*, ed. M. Keramidas, 563–70. Fairfax Station, VA: Interface Foundation.

Tsai, Tsung-han, and Jeff Gill. 2012. superdiag: A comprehensive test suite for Markov chain nonconvergence. *The Political Methodologist* 19:12–18.

Waterman, Richard W., and Kenneth J. Meier. 1998. Principal-agent models: An expansion? *Journal of Public Administration Research and Theory* 8(2):173–202.

Waterman, Richard W., Amelia Rouse, and Robert Wright. 1998. The venues of influence: A new theory of political control of the bureaucracy. *Journal of Public Administration Research and Theory* 8(1):13–38.

Whitford, Andrew B. 2002. Decentralization and political control of the bureaucracy. *Journal of Theoretical Politics* 14(2):167–93.

Witko, Christopher. 2006. PACs, issue context and congressional decision-making. *Political Research Quarterly* 59(2): 283–95.

———. 2011. Campaign contributions, access, and government contracting. *Journal of Public Administration Research and Theory* 21(4):761–78.

Wood, Dan B., and Richard W. Waterman. 1991. The dynamics of political control of the bureaucracy. *The American Political Science Review* 85(3):801–28.

Yackee, Susan Webb. 2006. Sweet talking the fourth branch: Influence of interest group comments on federal agency rule-making. *Journal of Public Administration Research and Theory* 16(1):103–24.