

2005

Elicited Priors for Bayesian Model Specifications in Political Science Research

Jeff Gill

University of California - Davis

Lee D. Walker

University of South Carolina - Columbia, walker23@mailbox.sc.edu

Follow this and additional works at: https://scholarcommons.sc.edu/poli_facpub



Part of the [Political Science Commons](#)

Publication Info

Published in *Journal of Politics*, Volume 67, Issue 3, 2005, pages 841-872.

This Article is brought to you by the Political Science, Department of at Scholar Commons. It has been accepted for inclusion in Faculty Publications by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.

Elicited Priors for Bayesian Model Specifications in Political Science Research

Jeff Gill

University of California–Davis

Lee D. Walker

University of Kentucky

We explain how to use elicited priors in Bayesian political science research. These are a form of prior information produced by previous knowledge from structured interviews with subjective area experts who have little or no concern for the statistical aspects of the project. The purpose is to introduce qualitative and area-specific information into an empirical model in a systematic and organized manner in order to produce parsimonious yet realistic implications. Currently, there is no work in political science that articulates elicited priors in a Bayesian specification. We demonstrate the value of the approach by applying elicited priors to a problem in judicial comparative politics using data and elicitations we collected in Nicaragua.

As quantitative political research becomes increasingly sophisticated, the more complex, but more capable, Bayesian approach is likely to grow in popularity. The Bayesian inferential engine is a coherent set of axioms that converts prior information to posterior evidence by conditioning on observed data. Thus, stipulating prior distributions for unknown quantities is a requirement, and this requirement has been a long-standing source of controversy. Bayesian statistics provide a number of ways to define prior information, and the strength of these prior assertions can vary considerably within the same inferential framework.

Recent Bayesian work in fields other than political science has exploited the *elicited prior* as a means of drawing information from subject-area experts with the goal of constructing a probability structure that reflects their specific qualitative knowledge, and perhaps experiential intuition, about the studied effects. These informed priors derive their name from the way in which the information is *elicited* from nonstatisticians who have a great deal of information about the substantive question but are not involved in the model construction process. Such experts can be physicians, policy-makers, theoretical economists, historians, previous study participants, outside experts, politicians, community leaders, and others.

Currently, there is almost no work in any social science field that advocates elicited priors in a Bayesian specification. Furthermore, nearly all published work on elicited priors exists in the medical trials¹ or engineering² literatures. This is unfortunate because elicited priors can be a means of systematically integrating qualitative and quantitative empirical work in political science, thus reaching across a traditional divide in the discipline. The promise of this approach is that it has the potential to tie together the seemingly antithetical research approaches of qualitative area studies with data-oriented work based on statistical methods, perhaps then mending a recent rift in political science (Bennett 2002; Kasza 2000). King, Keohane, and Verba (1994) argue forcefully for unification and rapprochement here, but, unlike the present study, they offer no specific methodology to accomplish this. In a response, the recent edited work by Brady and Collier (2004) ends up highlighting both the closeness between qualitative and quantitative work as well as the remaining cleft.

Early critics of the Bayesian paradigm (Fisher 1922; Neyman 1937; Pearson 1920a, 1920b) focused on the almost exclusive use of uniform (flat) priors at the time as a method for expressing prior ignorance or uncertainty. Their concern was the effect that this prior has with small samples (since large enough samples produce standard likelihood analysis results), and the fact that uniformness does not represent a genuine lack of information about a parameter. A great deal of Bayesian work in the middle of the twentieth century dwelt on the quixotic goal of finding an “objective” alternative (Jeffreys 1961; Lindley 1961; Savage 1972) to mitigate concern about arbitrarily interjecting subjective information through the prior. This effort proved to be misguided since all statistical models are subjective and a substantial *advantage* to the Bayesian choice is that previously known information can be directly and transparently included in the model specification.

The strongly informed approach taken in this article contrasts sharply with most existing Bayesian work in political science (Hill and Kriesi 2001; Jackman 2000a; Quinn, Martin, and Whitford 1999; Smith 1999; Western 1998), although one study (Jackman 2001) demonstrated a particular need for informed (but not elicited) priors. Recent efforts have instead focused on applying simulation tools from Bayesian statistics (i.e., Markov chain Monte Carlo) to solve previously intractable problems. This computationalist perspective mostly avoids the specification of deeply informed priors in favor of diffuse (very spread-out) forms (Jackman 2000b; Martin and Quinn 2002). Such priors have useful purposes (particularly in dealing with so-called nuisance parameters), but they do not fully

¹This literature elicits qualitative priors from clinicians as a means of incorporating such expertise into the calculations of posteriors and clinical trial stopping points (see Freedman and Spiegelhalter 1983 and Kadane 1986). Also, Garthwaite and Dickey (1988, 1992) and Ibrahim and Chen (2000) use elicited priors in this literature differently to make decisions about variable selection.

²The emphasis here is on risk assessment and management in large projects (Ayyub 2001; Blockley 1979; Cooke 1991; Dong and Wong 1986a, 1986b, 1986c; Furuta, Fu, and Yao 1985; Helmer 1967), and dates back to the *delphi method* developed at RAND in the 1950s.

exploit Bayesian capabilities. Without denigrating this approach, we provide here an alternative prior generation procedure at the opposite end of the information scale.

Several authors have argued strongly for the use of informed priors in the social sciences. Leamer (1972) states: "Arguments concerning the use of such prior information should appropriately address the question of how rather than whether prior information should be used" (1972, 1059). Beck and Jackman (1998) later note (in a non-Bayesian context) that the choice of smoothing parameters in generalized additive models should be "informed by prior beliefs" (1998, 608). Bartels (1996) uses prior information in pooling: "My analysis so far has emphasized that intelligent decisions about how to treat disparate observations must be based, in one way or another, upon prior beliefs about the statistical relevance of the available data" (1996, 917). Berk, Western, and Weiss (1995) observe that "... in complex studies where the prior information is based on clearly explained previous studies, the prior may find greater support among skeptical researchers than the model itself" (1995, 437). Thus we leave it to these authors and others to justify informedness in general, and we will concentrate on one particular form.

Background on Elicited Priors

Priors that are elicited from experts, also called assessors, have a variety of characterizations which are treated differently depending on the literature addressed. Kass and Greenhouse (1989) coined the phrase "community of priors" to describe the range of confidence or belief that equally qualified experts can have about the same effect, and Gill (2002, Chapter 5) categorizes elicited priors into four basic types. *Clinical Priors* are elicited from (generally easily captured) substantive experts who are taking part in the research project. *Skeptical Priors* are built with the assumption that the hypothesized effect does not actually exist and are typically operationalized through a probability function with mean zero. *Enthusiastic Priors* are the opposite of the sceptical prior and are usually built around the positions of partisan experts or advocates and generally assuming the existence of the hypothesized effect. *Reference Priors* are produced from expert opinion as a way to express informational uncertainty, but they are somewhat misguided in this context since the purpose of elicitation is to glean information that can be described formally. The general goal of all of these forms, however, is to translate a loose set of attitudes, opinions, and experiential knowledge into a probability statement that is useful statistically.

It is not necessary to have only one category of elicited prior, and it can be useful to contrast the posterior results obtained from divergent prior perspectives (see Western and Jackman 1994). Such an approach can include a formalized process of overcoming adversarial prior specifications in favor of priors more sympathetic to the research question through additional sampling (Lindley and Singpurwalla 1991), randomization strategies (Kass and Greenhouse 1989), or

scoring rules (Savage 1971). Moreover, the polling of “multiple expert opinion tends to perform well in the creation of estimates because it reflects the most up to date consensus” in the area of interest (Ascher 1978).

Particular Process Phases

The process of eliciting prior information from subject-matter experts can take many different basic forms, but the greater the formality of this process, the more *reliable* the results (i.e., giving consistent outcomes across repeated elicitations). Formalized approaches to gathering expert judgment have two distinct advantages over ad hoc recordings of opinion: more time and care is generally taken in creating questions for the elicitation process, and the steps are better documented. Spetzler and Staël von Holstein (1975) outline three general phases for the formal elicitation process.

The *deterministic phase* focuses on specifying explanatory variables in the model and possibly the assumed prior parametric form for their associated coefficients (Steffey 1992), determining the relevant data collection processes (Garthwaite and Dickey 1992), selecting the number of experts to query (Carlin et al. 1993), and planning how to evaluate the reliability of their contributions (Hogarth 1975). Some of this work is difficult: experts might need to be trained before elicitation (Winkler 1967), variable selection can be influenced by the difficulty of elicitation (Garthwaite and Dickey 1992), and cost estimates may be uncertain.

The *probabilistic phase* where experts are actually interviewed is generally the most difficult. There are three general approaches: assessors can be asked fixed value questions with probability responses (“P-methods”), fixed probability questions with value responses (“V-methods”), or more challengingly, questions to be answered on probability and value scales simultaneously (“PV-methods”; Spetzler and Staël von Holstein 1975, 347). P-methods determine interesting levels of explanatory variables (or perhaps a range of values in the case of interval measurement) and require the assessor to provide the probability of occurrence for levels of the outcome variable. Conversely, V-methods ask the more challenging question of determining explanatory variable levels associated with some given probability value. PV-methods are even more demanding because they require that the assessor pick cumulative distribution points and associated levels as a pair without prompting on either. Asking open-ended questions and coding the responses can be a very informative alternative method to these approaches, although unstructured answers are difficult to translate onto a probability metric. Such a free-form process, however, allows the assessors to introduce substantive issues that the investigator would not have otherwise thought to consider.

Finally, the *informational phase* is mostly mechanical, and it includes testing responses for internal consistency, calibrating these responses with reliable references, and sometimes weighting the assessors relative to each other. Consistency (no self-contradiction with an individual’s or a group’s answers) is an

important issue since assessors can differ in their familiarity with the studied effect, and not surprisingly, inexperienced assessors tend to give more internal inconsistencies, especially with continuous rather than discrete choices (Hogarth 1975; Winkler 1967). Unfortunately assessors sometimes give inconsistent answers in a probabilistic context such as specifying the probability of some subset of an event to have higher probability than the event itself (the conjunction fallacy). For example, one study found participants predicted the occurrence of more words ending in “-ing” rather than “-_n_” even though the former is included in the latter (Tversky and Kahneman 1983).

Constructing Elicited Priors

Since elicited priors are developed from substantive area experts rather than from the primary investigators themselves, the key challenge is translating verbal or written opinions into specific probability statements. This process ranges from informal assignments to detailed elicitation plans, and even regression analysis across multiple experts (Johnson and Albert 1999, Chapter 5). Commonly this process is automated so that assessors fill in values or select menu options on a terminal, see the subsequent effects to the prior, and then make adjustments to their original input.

A frequent strategy is to query experts about outcome variable quantiles for given (researcher-constructed) levels of specific explanatory variables (Kadane et al. 1980). For example, in one well-known study an emergency room physician is asked to estimate the survival probabilities of hypothetical patients with specified injury types, injury severity scores, trauma scores, and ages set by the researchers (Bedrick, Christensen, and Johnson 1997). This provides survival probabilities at various explanatory variable levels, so given an assumed distributional form for each of the priors, it is now possible to solve “backwards” for the needed prior parameters. More generally, assessors are first asked to give mean outcome variable responses or probability quantiles corresponding to a set of differing explanatory variable levels: $\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \dots, \tilde{\mathbf{X}}_m$. These $\tilde{\mathbf{X}}_i$ are called *design points* because they are vector-valued explanatory variable levels that are considered important or interesting by the researcher. The elicited outcome variable values now correspond to the “spread” of the full *design matrix* created by stacking these $\tilde{\mathbf{X}}_i$ vectors as if they were empirical cases. Prior distribution determination can be carried out via some accepted parametric model (Garthwaite and Dickey 1988), or conversely by scatterplot smoothing (Bedrick, Christensen, and Johnson 1997).

A related approach is to ask assessors to provide outcome-variable levels associated with specific cumulative probability levels (a V-method, the reverse of the operation just discussed). So the assessors provide values for the median, the 5% level, the 95% level, or other intuitive thresholds. The result can then be summarized nonparametrically, used as input into a pre determined parametric family of priors such as the normal/students- t (Al-Awadhi and Garthwaite 2001;

Garthwaite and Dickey 1988, 1992; Kadane 1980), or fit with piecewise linear approximations (Carlin et al. 1995).

Simple Elicitation Using Linear Regression

Consider a basic example where the analyst asks an expert for predictions on an expected outcome for an interval-measured event. The (V-method) question asked is what would be an expected low value in the form of a .25 quantile (x_1) and an expected high value in the form of a .75 quantile (x_2). These values then help specify a normal distribution for this event (other prior forms can be specified in similar fashion). This is a Bayesian process that should appeal to frequentists since the assumed prior distribution matches standard asymptotic theory.

The two supplied quantile values, x_1 and x_2 , corresponding to $z_1 = .25$ and $z_2 = .75$, exactly specify the shape of a normal PDF since there are two unknown parameters and two equations:

$$z_1 = \frac{x_1 - \alpha}{\beta} \quad z_2 = \frac{x_2 - \alpha}{\beta}, \quad (1)$$

where α and β are the mean and standard deviation parameters of the normal form: $f(x|\alpha, \beta) = (2\pi\beta^2)^{-\frac{1}{2}}\exp[-(x - \alpha)^2/2\beta^2]$. Therefore when we solve for α and β we have a fully defined prior distribution from the elicitation.

One expert is often insufficient so we now query experts 1, 2, . . . , J , producing an overspecified series of equations since there are $J \times 2$ equations and only two unknowns (Spiegelhalter, Freedman, and Parmar [1994] use $J = 10$ for example). We assume for now that these experts are *exchangeable*, meaning that they all provide equal quality elicitations. Actually, given the cost of interviewing, we are much more likely to ask each expert for more than just two quantiles, and it is always helpful to have more assessed points if they are deemed reliable. So each assessor is asked to give five quantile values at $m = [.01, .25, .5, .75, .99]$ corresponding to standard normal points z_m . Now (1) can be reexpressed for the quantile level m given by assessor j : $x_{jm} = \alpha + \beta z_{jm}$. Therefore the total amount of expert-elicited information constitutes the following overspecification ($J \times 5$ equations and 2 unknowns) of a normal distribution:

$$\begin{array}{lllll} x_{11} = \alpha + \beta z_{11} & x_{21} = \alpha + \beta z_{21} & \cdots & x_{(J-1)1} = \alpha + \beta z_{(J-1)1} & x_{J1} = \alpha + \beta z_{J1} \\ x_{12} = \alpha + \beta z_{12} & x_{22} = \alpha + \beta z_{22} & \cdots & x_{(J-1)2} = \alpha + \beta z_{(J-1)2} & x_{J2} = \alpha + \beta z_{J2} \\ x_{13} = \alpha + \beta z_{13} & x_{23} = \alpha + \beta z_{23} & \cdots & x_{(J-1)3} = \alpha + \beta z_{(J-1)3} & x_{J3} = \alpha + \beta z_{J3} \\ x_{14} = \alpha + \beta z_{14} & x_{24} = \alpha + \beta z_{24} & \cdots & x_{(J-1)4} = \alpha + \beta z_{(J-1)4} & x_{J4} = \alpha + \beta z_{J4} \\ x_{15} = \alpha + \beta z_{15} & x_{25} = \alpha + \beta z_{25} & \cdots & x_{(J-1)5} = \alpha + \beta z_{(J-1)5} & x_{J5} = \alpha + \beta z_{J5} \end{array}$$

The solution suggested by this setup is to run a simple bivariate linear regression with α as the intercept and β as the slope. Critically, one must check for logical

inconsistencies both in consistent quantile values for each assessor (see Lindley, Tversky, and Brown [1979]), and in mathematical constraints such as ensuring that the estimated coefficient for β remains positive (if substantively required) since the basic linear model imposes no such restriction (Raiffa and Schlaifer 1961).

Suppose we are interested in eliciting a prior distribution for expected campaign contributions received by major-party candidates in an upcoming election for contested U.S. Senate seats as part of a larger Bayesian specification. Since this cycle has not yet occurred we cannot simply obtain these data from the Federal Election Commission. An alternative strategy is to elicit opinions from election experts on what contribution levels they might expect to see. This is a good example of the value of elicitation since many people expert in campaign contributions are practitioners or analysts outside of academic political science.

Eight experts are queried for quantiles at levels $m = [.1, .5, .9]$, and they provide the following values reflecting the national range of expected total intake by Senate candidates (in thousands):

| | | | | | | | |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $x_{11} = 400$ | $x_{21} = 150$ | $x_{31} = 300$ | $x_{41} = 250$ | $x_{51} = 450$ | $x_{61} = 100$ | $x_{71} = 500$ | $x_{81} = 300$ |
| $x_{12} = 2500$ | $x_{22} = 1000$ | $x_{32} = 900$ | $x_{42} = 1200$ | $x_{52} = 1800$ | $x_{62} = 1000$ | $x_{72} = 2100$ | $x_{82} = 1200$ |
| $x_{13} = 4000$ | $x_{23} = 2500$ | $x_{33} = 1800$ | $x_{43} = 2000$ | $x_{53} = 3000$ | $x_{63} = 2500$ | $x_{73} = 4200$ | $x_{83} = 2000$ |

(recall, for instance, that x_{83} indicates expert eight's third quantile). Since none of the experts have supplied quantile values out of logical order, these results are consistent. Now we have sufficient "data" to regress x on z , obtaining intercept and slope values: $\alpha = 276$, $\beta = 2368$. So the normal prior median value (also the mean) is obtained by $x_{.50} = 276 + 2368(.5) = 1460$, and the normal prior standard deviation is obtained by $x_{.84} - x_{.50} = 276 + 2368(.84) - 1460 = 805$. Thus we have complete information to specify the elicited normal prior for this explanatory variable coefficient in an encompassing Bayesian model.

Variance Components Elicitation

A pervasive problem with direct quantile elicitation procedures is that assessors tend to misjudge the occurrence of unusual values because it is more difficult to visualize and estimate tail behavior than to estimate means or medians. When nonstatistical assessors are asked to estimate spread by providing high probability coverage intervals such as at 99%, then there is an inclination to perceive this as near-certainty coverage and overstate the bounds (Hora, Hora, and Dodd 1992). Conversely, in other popular settings people tend to think of rare events in the tails of distributions as more likely than they really are. Casinos and state lotteries play on this weakness by overemphasizing winning events relative to many other losses. O'Hagan (1998) suggests improving elicited estimates of spread by separately requiring assessors to consider two types of uncertainty: uncertainty about an estimate relative to an assumed known summary statistic, and the uncertainty of this summary. First the assessor gives a (modal) point esti-

mate for the explanatory variable coefficient: τ . Then they are asked: “given your estimate of τ , what is the middle 50% probability interval around τ ?” Assessors need to understand that this is the interval that contains the middle half of the expected values. So this V-method specifies a density estimate centered at the assessor’s modal point, and if the form of the distribution is assumed or known, then the exact value for the variance can be backed out mathematically. Often the assumed distribution is normal or students- t , but a log-normal form can be used if the assessor specifies a right-skewed interval.

O’Hagan (1998) actually prefers asking for the middle 66% of the density (he calls this the “two-to-one interval” since the middle coverage is twice that of the combined tails). If a normal prior is assumed then this interval quickly yields a value for the standard deviation since it covers approximately two of them (it should actually be multiplied by $\frac{68}{66}$ but analysts typically do not worry about the difference). Once the assessor gives this interval, the researcher calculates the implied variance and shows the assessor credible intervals (Bayesian confidence intervals) at familiar $(1 - \alpha)$ -levels, such as 50% or 99%, so that the assessor can see the general implications of their assigned spread. If these are deemed to be too large or too small, then the process is repeated.

Suppose that the purpose of elicitation is to obtain prior distributions for unknown values τ_i across n cases, with unknown total $T = \sum_{i=1}^n \tau_i$. The assessor first provides point estimates for each case: x_1, x_1, \dots, x_n , so that the estimated total is given by $x_T = \sum_{i=1}^n x_i$. These are useful values but it is still necessary to get a measure of uncertainty in order to produce a variance for the full elicited prior distribution.

The individual deviance of the i th estimate from its true value can be rewritten algebraically:

$$\tau_i - x_i = \left(\tau_i - \frac{x_i}{x_T} T \right) + \frac{x_i}{x_T} (T - x_T). \quad (2)$$

The first quantity on the right-hand-side of (2) is the deviance of τ_i from an estimate that would be provided if we knew T for a fact:

$$E[\tau_i | T] = \frac{x_i}{x_T} T, \quad (3)$$

which can be considered as between-case deviance. The second quantity on the right-hand-side of (2) is the weighted deviation of T , i.e., uncertainty about the true total. The expected value form (3) helps us obtain the variance of τ_i :

$$\begin{aligned} \text{Var}(\tau_i) &= E[\text{Var}(\tau_i | T)] + \text{Var}(E[\tau_i | T]) \\ &= E \left[\tau_i - \frac{x_i}{x_T} T \right]^2 + \left(E[E[\tau_i | T]^2] - (E[E[\tau_i | T]])^2 \right) \\ &= E \left[\tau_i - \frac{x_i}{x_T} T \right]^2 + \left(\frac{x_i}{x_T} \right)^2 \text{Var}(T), \end{aligned} \quad (4)$$

which shows the general form of the two variance components. A more natural form for elicitation is achieved by dividing both sides of this equation by x_i^2 :

$$\text{Var}\left(\frac{\tau_i}{x_i}\right) = \text{Var}\left(\frac{\tau_i}{x_i} - \frac{T}{x_T}\right) + \text{Var}\left(\frac{T}{x_T}\right). \quad (5)$$

Now assessors can be queried about the middle spread around the two quantities separately. First, they are asked to give an estimate of middle spread around $\frac{T}{x_T}$, assuming accuracy of the sum x_T as an estimate of T . Second, they are then asked for the middle spread around each $\frac{\tau_i}{x_i}$, temporarily assuming again that $\frac{T}{x_T} = 1$ so there is no second component to the variance to consider. Once the individual means and variances are elicited, these $\frac{x_i}{x_T}$ values can be plugged into an assumed distribution defined over $[0:1]$ (since they are proportions) to create a complete prior distribution specification. Two direct forms are the normal CDF (see the example below) and the beta distribution. If $\frac{\tau_i}{T}$ is assumed to be distributed beta, then we can readily solve for the parameters with the definitions of the beta distribution mean and variance: $E\left[\frac{\tau_i}{T}\right] = \frac{\alpha}{\alpha + \beta}$, $\text{Var}\left(\frac{\tau_i}{T}\right) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$.

This approach is illustrated with the following contrived example. An expert on minority electoral participation is asked to estimate upcoming Hispanic turnout for n precincts in a given district: $\tau_1, \tau_2, \dots, \tau_n$, with total Hispanic turnout in the district equal to T . The expert first gives estimates x_1, x_2, \dots, x_n for each precinct which produces a district turnout estimate of T by summing, x_T . Of course this result is useful but it does not yet give the variance information necessary to build a prior using an assumed normal distribution.

The expert is then asked to provide the two-to-one interval for $\frac{T}{x_T}$, giving $[.7:1.3]$: they believe that the summed estimate of Hispanic turnout is correct to plus or minus 30% with probability .66 (from the two-to-one interval). To test the expert's conviction about this variance, the value $\sigma_T = .3$ is plugged into the appropriate normal CDF at levels to give the credible interval summaries:

$$\begin{aligned} 50\%CI &= [\Phi_{\mu=1, \sigma=.3}(.25) : \Phi_{\mu=1, \sigma=.3}(.75)] & 99\%CI &= [\Phi_{\mu=1, \sigma=.3}(.005) : \Phi_{\mu=1, \sigma=.3}(.995)] \\ &= [.798 : 1.202] & &= [.227 : 1.773] \end{aligned}$$

which are then read back to the expert. If they agree that these are reasonable summaries then the variance is set to $\sigma_T^2 = (.3)^2 = .09$ and there is no need to iterate. Next the expert is asked to repeat this process for each of the x_i estimates under the temporary assumption that $x_T = T$ (that the estimate of the total above is correct). This "certainty" means that the right-hand-side of (5) reduces to the variance of $\frac{\tau_i}{x_i}$ and the expert can perform the same interval process as was done with $\frac{T}{x_T}$ for each of the n precincts. Suppose that two-to-one interval for the estimate of Hispanic turnout at the first precinct ($x_1 = .2$) is given as $[.5:1.5]$, meaning that the estimate is believed to be correct to plus or minus 50% with probability .66. This implies a variance of $\sigma_1 = (.5)^2 = .25$, and we will assume that the subsequent 50% and 99% credible interval summaries are approved by the expert. Therefore the total elicited variance for the first precinct is given by (5) where

x_1^2 is moved back to the right-hand-side: $\text{Var}(\tau_1) = x_1^2(.25 + .09) = .0136$. Notice, incidentally, the dependency here on the modal estimate x_1 .

Predictive Modal Elicitation

If the outcome variable of interest is distributed Bernoulli or binomial (coup/no-coup, vote/not-vote, etc.), then it is usually straightforward to query experts *directly* for prior probabilities. Apparently for psychological reasons probabilities of binary or summed binary outcomes are intuitively easy to visualize in substantive terms (Cosmides and Tooby 1996). Using a beta conjugate prior Chaloner and Duncan (1983, 1987) develop the predictive modal (PM) elicitation algorithm (see also Gavasar 1988). Conjugacy in a Bayesian model means that the resulting posterior will also be in the same parametric family as the prior (a beta distributional form in this case).

The general steps are to first fix a hypothetical total number of Bernoulli trials, ask the assessor to specify the most likely number of successes as well as reasonable bounds on the uncertainty, and then work these values backward into the beta-binomial parametric setup to get the implied beta prior distribution parameters. Then the assessor is shown the implications of their stipulated values on the shape of the beta prior. If the assessor finds the results to be substantively unreasonable, then adjustments are made in the deterministic phase and the process is repeated.

More formally, select a fixed number of trials for a hypothetical experiment; $n = 20$ is recommended in general situations, but the choice can depend on the complexity of the question. The assessors are then asked to give the prior predictive modal value for this n : the most likely number of successes out of n trials, m . This value can range from $m = 1$ to $m = n$, but if assessors are allowed to pick $m = 0$, then more involved approaches are required. Since the data, X_1 ,

X_2, \dots, X_n , are distributed iid Bernoulli, then we know that $Y = \sum_{i=1}^n X_i$ is distributed binomial(n, p). Our primary interest lies in the posterior distribution of the probability of occurrence of some political event of interest (bill passes/fails, treaty/no-treaty, etc.), which is p in the binomial PMF. The PM method first assumes a beta distribution prior for p with unknown parameters: $p \sim \text{beta}(A, B)$. In addition, the two parameters of this beta distribution are both required here to be greater than one ($A, B > 1$) in order to produce a unimodal form (a frequent but not necessary assumption; see the extension in Gavasar 1988).

The joint distribution of y and p given values of A and B is:

$$\begin{aligned} f(y, p|A, B) &= f(y|p)f(p) = \left[\binom{n}{y} p^y (1-p)^{n-y} \right] \times \left[\frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} p^{A-1} (1-p)^{B-1} \right] \\ &= \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)} p^{y+A-1} (1-p)^{n-y+B-1}. \end{aligned} \quad (6)$$

The marginal distribution for y is obtained by integrating out p :

$$\begin{aligned} f(y|A, B) &= \int_0^1 \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)} p^{y+A-1}(1-p)^{n-y+B-1} dp \\ &= \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)} \frac{\Gamma(y+A)\Gamma(n-y+B)}{\Gamma(n+A+B)}. \end{aligned} \quad (7)$$

Since m is the mode of this distribution, then $f(y = m|A, B)$ is the maximum value obtainable for the function $f(y|A, B)$. Recall that y can only take on discrete values so it must be true that $f(y = m - 1|A, B) < f(y = m|A, B)$ and $f(y = m + 1|A, B) < f(y = m|A, B)$. Chaloner and Duncan then calculate the following two ratios using (7):

$$\begin{aligned} d_\ell &= \frac{f(y = m - 1|A, B)}{f(y = m|A, B)} = \frac{(n-m)(m+A)}{(m+1)(n-m+B-1)} \\ d_r &= \frac{f(y = m + 1|A, B)}{f(y = m|A, B)} = \frac{m(n-m+B)}{(n-m+1)(m+A-1)}, \end{aligned} \quad (8)$$

which are both necessarily bounded by (0:1). Why is this useful? Once the assessor has identified m for the researcher, then the prior parameters (A, B) must be constrained to lie in a cone originating at $[1, 1]$ in the A, B plane as shown by the solid lines in Figure 1. This cone is determined because the equations in (8) define linear limits in two-space from the same starting point. Points within the cone represent Cartesian distance from a uniform prior (a benchmark) since the origin of the cone specifies a beta(1, 1), which is a uniform PDF.

Of course we do not yet have a complete answer since there are an infinite number of (A, B) pairs that could be selected and still remain inside the cone. Now the assessor is told to think about *spread* around the mode and is shown a histogram for binomial($n, m/n$) (either computerized via interactive screens or the researcher must have $n = 20$ plots ready to show). The question is:

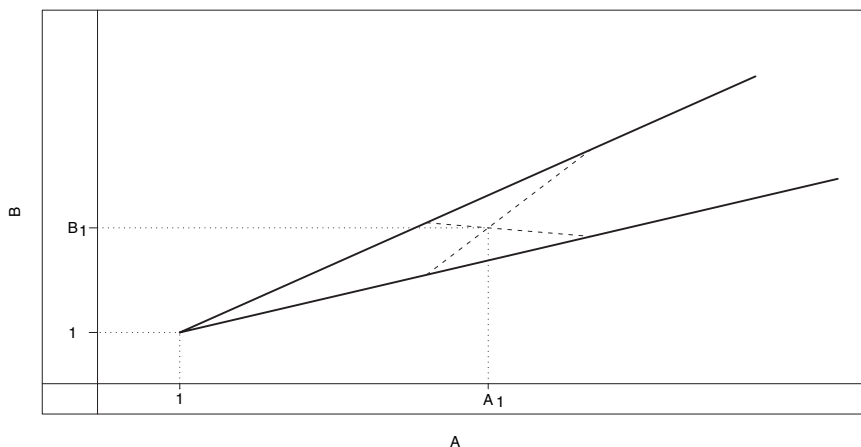
If we were to go one unit up (and down), how much do you think the probability of occurrence would decrease?

Armed with these two values (up and down) we can calculate values of d_ℓ and d_r directly, so the equations in (8) define line segments for values of A and B which will necessarily be bounded by the cone. The point of intersection is the (A, B) pair that satisfies both the one unit up restriction *and* the one unit down restriction, provided that

$$d_\ell d_r > \frac{m(n-m)}{(m+1)(n-m+1)} \quad (9)$$

holds (the direction of inequality is wrong in Chaloner and Duncan [1983]). If it does not, then the assessor is asked to provide new values of $f(y = m - 1|A, B)$ and $f(y = m + 1|A, B)$. These two line segments and their intersection are also shown in Figure 1. Call this point of intersection (A_1, B_1) , calculate a new beta mode with these values:

FIGURE 1
Finding the Beta Parameters



$$m_1 = \frac{A_1 - 1}{A_1 + B_1 - 2}, \quad (10)$$

and then display it to the assessor with the middle 50% of the density. The assessor is then asked if this interval is too small, too large, or just right. The interval is adjusted according to the following:

“too small” $\Rightarrow h = -1$, “too large” $\Rightarrow h = +1$, “just right” $\Rightarrow h = 0$,

where h is inserted into: $A_2 = 1 + 2^h(A_1 - 1)$, $B_2 = 1 + 2^h(B_1 - 1)$,

for adjusted parameter values (although the original modal value is preserved). This process repeats until the assessor is satisfied with the interval ($h = 0$). Note also that in repeatedly respecifying their probability drop-off, assessors see the sensitivity of the interval to changes in their values.

This is obviously a vast improvement over asking subject-matter experts to give beta or Dirichlet distribution parameters directly, as was done by earlier researchers (Bunn 1978, 1979). Carlin et al. (1993), Kadane et al. (1980), Chaloner et al. (1993), and others have developed PM technology with interactive graphical displays that let the assessors iteratively manipulate a graphical image of the prior on the screen and change it dynamically with a computer mouse or other instruments until they feel that the resulting interval reflects their beliefs.³

³Lara Wolfson wrote and freely distributes the package *elicit-normlin* based on Kadane (1980) which can be downloaded at <http://lib.stat.cmu.edu/general/>. Ram Gopalan's program for eliciting Dirichlet process priors, *elicit-diric*, is also available at the same site.

As a simple example, suppose we are interested in eliciting a prior distribution for the probability of a strike given a major labor dispute in 20 OECD countries over the next year. Given $n = 20$ an assessor returns $m = 5$, and is then shown a histogram of the binomial(20, 5/20) distribution. The assessor now indicates that the one unit up and down probability change is 1/50, so

$$d_l = .98 = \frac{(15)(5+A)}{(6)(14+B)} \quad \text{and} \quad d_r = .98 = \frac{5(15+B)}{(16)(4+A)} \quad (\text{these values are acceptable})$$

since $d_l d_r = .9604 > \frac{m(n-m)}{(m+1)(n-m+1)} = .7813$. Solving the equations produces

$A_1 = 1.4$, and $B_1 = 2.327$, which gives a modal value for the probability of a strike given a labor dispute of $\hat{p} = .232$ with the middle of 50% of the density [.193:.537]. Our assessor believes that this interval is too large, so we set $h = +1$ and produce $A_2 = 1 + 2^h(1.4 - 1) = 1.8$, and $B_2 = 1 + 2^h(2.327 - 1) = 3.654$, which gives the middle of 50% of the density as [.184:.456]. This is still considered too large by the assessor so the process is repeated with $h = +1$ now providing [.182:.386]. This interval is noticeably smaller than the first and is acceptable to the assessor, meaning that the elicited prior distribution for the probability of a strike given a major labor dispute has the distribution beta(2.6, 6.308).

Prior Elicitation for the Normal-Linear Model

The normal-linear regression model is ubiquitous in political science and has been a methodological workhorse for at least 50 years. Bayesian applications have also appeared recently (Bartels 1996; Francisco 1996; Gerber and Jackson 1993; Western and Jackman 1994), including one economic policy application with elicited priors (Leamer 1992). Start with the well-known basic multiple linear regression model conforming to the Gauss-Markov assumptions, and define terms conventionally: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where \mathbf{X} is an $n \times k$, rank k matrix of explanatory variables with a leading column of ones for the constant, $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients to be estimated, \mathbf{y} is an $n \times 1$ vector of outcome variable values, and $\boldsymbol{\varepsilon}$ is a $n \times 1$ vector of errors distributed $N(0, \sigma^2 \mathbf{I})$ for a constant σ^2 . The classical Bayesian approach to diffuse priors for this model specifies: $p(\boldsymbol{\beta}) \propto c$ over $(-\infty, \infty)$ for an arbitrary constant c , and $p(\sigma^2) = \frac{1}{\sigma}$ over $(0, \infty)$ (Tiao and Zellner 1964, 220). This “noninformed” approach produces the marginal posteriors:

$$(\boldsymbol{\beta} - \mathbf{b})|\mathbf{X}, \mathbf{y} \sim t_{n-k} \quad (\text{provided that the covariance matrix, } \mathbf{R} = \frac{(n-k)\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}}{n-k-2},$$

is positive definite), and $\sigma^2|\mathbf{X}, \mathbf{y} \sim IG(a, b)$. The latter is an inverse gamma distribution with $a = (n - k - 1)/2$ and $b = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/2$.

Instead of this vague approach just described, we want to elicit opinions for priors on $\boldsymbol{\beta}$ from our experts (we can retain the noninformed approach for σ^2 or we can elicit for it as well). Kadane et al. (1980) suggest the following approach (see also the similar but more recent summary in Kadane and Wolfson 1998).

First establish m design points of the explanatory variable vector: $\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \dots, \tilde{\mathbf{X}}_m$, where again these represent interesting cases or values spanning the range of the k variables. It is important that these values be chosen such that stacking the vectors into a $m \times k$ matrix, $\tilde{\mathbf{X}}$, gives a positive definite matrix $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ (it is quite possible to do otherwise!).

Next the assessors are asked to study each of the $\tilde{\mathbf{X}}_i$ row vectors and produce $\mathbf{y}_{.50}$, a vector of outcome variable medians whose elements correspond to the design cases. These values then represent “typical” responses to the hypothesized design points specified in the $\tilde{\mathbf{X}}$. Therefore an elicited prior point estimate for β is given by: $\mathbf{b}_{.50} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{y}_{.50}$, which is quite intuitive.

We need, however, to get a prior *distribution* for β not just a prior point estimate. Assume, given the discussion above, that this distribution is students- t around $\mathbf{b}_{.50}$ with greater than two degrees of freedom (so we do not have to worry about the existence of the first two moments). Thus we are actually specifying a somewhat conservative prior since large data size under weak regularity conditions leads to Bayesian posterior normality of linear model coefficients, and t -distributed forms with smaller data size (Berger 1985, 224; Lindley and Smith 1972). Unfortunately there is no direct guidance about setting the degrees of freedom for this t -distribution since the m value was established arbitrarily by the researchers, and it is not generally helpful to elicit a degrees-of-freedom parameter directly from subject matter experts. Obtaining it from the data here is not helpful either because the elicitation process is supposed to take place before conditioning on the observations.

To solve this annoying problem Kadane et al. (1980) suggest a further continuation of the described procedure. After eliciting $\mathbf{y}_{.50}$ for each $\tilde{\mathbf{X}}_i$, also elicit $\mathbf{y}_{.75}$ by asking for the median point above the median point just provided. Elicit now according to this scheme two more times in the same direction to obtain $\mathbf{y}_{.875}$, and $\mathbf{y}_{.9375}$. Then for each of the m assessments calculate the ratio:

$$a(\tilde{\mathbf{X}}) = (\mathbf{y}_{.9375} - \mathbf{y}_{.50})/(\mathbf{y}_{.75} - \mathbf{y}_{.50}), \tag{11}$$

where the subtraction makes the numerator and denominator independent of the center, and the ratio produced is now independent of the spread described. This ratio uniquely describes tail behavior for some t -distribution because it is the relative “drop-off” in quantiles. Kadane et al. tabulate degrees of freedom against values of this ratio for simple lookup, a subset of which is given below:

| | | | | | | | | | |
|-------------------------|------|------|------|------|------|------|------|----------|------|
| df | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 |
| $a(\tilde{\mathbf{X}})$ | 2.76 | 2.62 | 2.53 | 2.48 | 2.45 | 2.42 | 2.40 | 2.39 | 2.37 |
| df | 14 | 16 | 18 | 20 | 30 | 40 | 60 | ∞ | |
| $a(\tilde{\mathbf{X}})$ | 2.36 | 2.35 | 2.34 | 2.33 | 2.31 | 2.31 | 2.30 | 2.27. | |

Values greater than 2.76 indicate that the researcher should instruct the assessor to reevaluate their responses, and values less than 2.27 imply that a standard

normal prior centered at $\mathbf{b}_{.50}$ can be used. For the example of prior elicitation with the linear model we will defer until the substantive example below which follows the same logic but gives an extension to nonlinear forms.

Reliability Analysis of Elicited Priors

Subjective probability assessments are a measure of an individual's "degree of belief" and are thus variable over time as well as situationally. Because elicited priors are generated by humans through a process that partially distances them from the actual mathematical construction of the prior distribution, it is critical to check the validity of the resulting form. The four major concerns are: *coherence*, *consistency*, *calibration*, and *bias*.

Elicited Prior Coherence

If P-methods are used to elicit prior information from assessors then a primary concern is *coherence*, ensuring that the resulting probability statements produce valid probability functions. This means that the probabilities given cannot lead to a violation of the standard Kolmogorov probability axioms. How can particular answers lead to violations here? Suppose that the assessor gives the probability of the Republican candidate winning at .60, and the probability of the Democratic candidate winning at .50. Since these are pairwise disjoint events, then the probability of their union (the probability of a major party candidate winning) is the sum of their probabilities which is 1.10. This is clearly an oversimplified example with an easily detectable problem, but suppose that the assessor was providing probabilities for many more pairwise disjoint events, something like A_1, A_2, \dots, A_{20} (perhaps the predicted electoral fortunes of Italian political parties). In this case the assessor would have to mentally track the interim sum of the probabilities, which may be difficult.

Consider an example modified from Lindley, Tversky, and Brown (1979) that defines for one country the events: L for losing a war, T for settling a war by treaty, O concluding a war by another means (victory, armistice, temporary cessation), and W^c for no war. The expert elicitations are:

$$p(L) = .33 \quad p(T) = .27 \quad p(O) = .23 \quad p(W^c) = .12$$

$$p(L|W) = .41 \quad p(T|W) = .31 \quad p(O|W) = .28.$$

This is incoherent since $p(L) + p(T) + p(O) + p(W^c) = .95$, even though $p(L|W) + p(T|W) + p(O|W) = 1.00$. More subtly, the conditional probability ratios do not coincide with the unconditional probability ratios: $p(L|W)/p(T|W) = 1.32$, but $p(L)/p(T) = 1.22$, which is a logical disagreement since L and T are both subsets of the event W . Also, when interval estimates are substituted for point estimates additional incoherences can appear, such as probability regions with logical

conflicts. For example, if part of the region for $P(L|W)$ existed outside of the region for $P(W)$, then it would imply the possibility of losing a war without going to war. Well-designed elicitation studies therefore evaluate probability function consequences “on the spot” (usually via interactive computer screens) and ask the assessor to respecify until violations are removed.

V-methods are generally safer from incoherence because researchers control the probability levels, and ask for values. The only necessary caution stems from the possibility of substantively illogical directional responses. For instance, the following paired level responses are incoherent:

| Question | Response |
|----------------------------------------------------------------------------------------------------------------------|----------|
| If there is 60% probability that the war lasts an additional two years, how many total casualties do you anticipate? | 5,000 |
| If there is 20% probability that the war lasts an additional two years, how many total casualties do you anticipate? | 8,000 |

The incoherency therefore stems from the notion that prolongation of the war reduces casualties (actually, there *may* be circumstances where this could occur but researchers can judge the reasonableness of such claims). We could also see this type of incoherency effect with credible intervals if the assessed values lead to intervals where lower α -levels provide narrower prior coverage, and vice-versa. For example, if for the same problem a 99% credible interval indicates a smaller region than a 95% credible interval, then there is an obvious incoherency.

Elicited Prior Consistency

Consistency means that the assessor applies the same subjective personal criteria across events. Therefore consistency is a property *within* each assessor that is violated when the individual produces probabilities or levels that are not directly comparable. Consider the infamous “Russian judge” stereotype for grading figure skaters where the defining characteristic is that he or she gives low grades for *all* skaters relative to the other judges. This is not considered a problem since figure skating contests are won by relative scores and as long as the Russian judge is *consistent* in their harshness, the contest is still viewed as “fair.” Consistency is then a criteria independent of the “truth” since it applies to within-expert concordance.

Unfortunately, consistency measured between assessors is more complex. Whenever information is acquired from multiple assessors, the resulting prior is a combination of these elicitations. Probability or level statements can be averaged, or averaged with weighting. If the variable is discrete and it is essential to preserve this discreteness in the prior, then the elicitations can be considered “votes” and the prior is the histogram of the resulting values (Carlin et al. 1993). Since a histogram is actually a density estimate, prior uncertainty can be obtained

parametrically or nonparametrically from the variance of this histogram. Other strategies include applying decision-theoretic rules (minimax regret, etc.) or various forms of pooling (Kadane 1986). When combining elicitations across assessors in these ways, the general definition of consistency changes to include directional consistency so that the aggregation scheme selected does not lead to illogical statements just because assessors differ in their directions. Winkler (1967) recommends comparing the implied cumulative density functions for each assessor as a means of checking consistency between assessors. Press (1989, Chapter 4) uses several graphic approaches to judging consistency. The basic worry is that techniques like averaging can “hide” features of the individual elicitations if not checked.

Elicited Prior Calibration

Seidenfeld (1985) formalizes criteria for *calibrated* subjective interpretation of probability statements as a means of verifying good elicitation:

- **Coherence.** Assessors have “belief-states” that are modeled by well-behaved conditional probability statements.
- **Total Evidence.** Assessors include all relevant information known at the time of elicitation, including background knowledge as well as the information directly queried by the researcher.
- **Conditionalization.** Assessors update their knowledge when new evidence is observed in the same way that Bayesian prior information is updated by conditioning on observed data provided through the likelihood function.

These criteria imply that the process of elicitation is inherently Bayesian in nature, even before the formal calculation of the posterior (for a counter-argument, see Kahneman and Tversky 1982). Essentially Seidenfeld is saying that the central axioms of Bayesian inference (probability statements as a foundation, inclusion of all pertinent prior information, and updating based on collected observations) define a calibrated elicitation process. Note also that Seidenfeld’s definition overlaps with the structure provided here since, for him, coherence is a subcategory of calibration.

Calibration is also defined in terms of inter-rater consistency across elicitations (Smith et al. 1978). Lindley, Tversky, and Brown (1979) point out that calibration is a characteristic of multiple assessments compared with past accuracy. To use their example, a well-calibrated meteorologist is one whose record of prediction is correct on average: rain occurs two-thirds of the time that they make a two-third prediction of rain. That is, the proportion of correct statements with the same probability is equal to the actual proportion of occurrence (although variances could differ as well, giving a different view of reliability). So calibration differs from consistency since it implies a direct comparison with subsequently observed events that are either confirming or disconfirming.

Bias Issues with Elicited Priors

Expert bias⁴ is a problem because it can harm the quality of the results by altering inferences in a direction away from the target population values, even with reasonably large sample size (Crosby 1980). Meyer and Booker (2001) identify two primary sources of expert bias. *Motivational bias* occurs when the expert responds to social group pressure within the project or research environment generally. Under this scenario the assessor alters responses in an effort to please others involved in the work or by a wish to positively affect the outcome of the study. This is often easy to diagnose since the individual's responses will deviate strongly from other cases, and such a suspect assessment can be benchmarked with clinical or enthusiastic priors, since it will tend to resemble them. *Cognitive bias* comes from: inconsistency (discussed above), "anchoring," memory limitations, underestimation of uncertainty, and confusing probability levels with personal utilities (see Kadane and Winkler 1988 on this latter point). Anchoring is a psychological phenomenon in assessors whereby they fix an early response as a reference point for subsequent responses, whether this is appropriate or not. That is, assessors seek internal consistency and ease of answering by anchoring statements against one particular statement that is important or easy to remember. Bias from memory limitations occurs since some events are easier to recall because they were: notably tragic, more recent, more concrete, or more personal. Thus bias can occur if the questions elicit information with a mixture of saliency (Cooke 1991). Bias from underestimation of uncertainty is deeply personal and stems from the observation that people tend to dislike chaos and uncertainty in their lives and therefore overestimate ordering effects when observed. So future events seemingly occur with greater certainty to produce overconfident predictions. These types of biases are usually combated with more sophisticated question wording which queries specific subject matter with alternate phrasing or circumstances (this is often done in exactly the same way that survey researchers have prescribed methods for increasing *construct validity*—see Groves 1987, Judd and McClelland 1984).

Bias also comes from a (familiar) third source: sample selection. The assessor or group of assessors picked for the study could have some systematic bias about probabilities of occurrence based on their backgrounds. For instance, querying Defense Department experts and State Department experts about the probability of a war can lead to radically different answers due to their respective professional orientations (trained to fight wars versus trained to avoid wars). Thus picking assessors is an important part of an elicitation study and the researcher should be overt about the implemented selection process. Since the definition of

⁴Expert bias differs from statistical bias ($\theta - \hat{\theta}$) in the sense that it results from not accounting for the mental model that the assessor uses in responding to the questions. Humans observe reality through their own sense of context and therefore it is essential to understand the general perspective of the elicitation group. See Meyer and Booker (2001, 38–44) for an extended discussion.

what constitutes an expert is entirely field-specific, general guidance on expert selection is impractical. However, many areas in political science have well-defined conventions about who has expert knowledge: senior jurists, legislators, specialized policy analysts, diplomats, regulators, agency-heads, party professionals, etc.

Elicitor selection bias can often be detected by picking a broad range of assessors and contrasting their responses to key questions. If individuals with similar backgrounds cluster around certain responses far afield of others, then this is an indication that this group has distinct perspectives and therefore possible biases. Obviously, the larger the group of assessors queried, the lower the possibility that their aggregate view is biased.

Meyer and Booker (2001) give some specific advice for reducing potential bias in a given study:

- Anticipate potential biases and test for their existence through inter- and intra-assessor comparisons.
- Design, and possibly redesign, the study to minimize anticipated or observed biases.
- Minimize bias through pretraining of assessors.
- Monitor the elicitation process and check the full set of elicitations.

In other words: be careful! This involves thorough consideration of assessors who are involved in the study, review of assessor professional background, consideration of the effect of question wording and order, as well as including training and monitoring as part of the study. In the next section we return to these issues in our application. While much of the guidance in the literature centers on clinical trials in medicine where addressing some of these issues is not a great challenge, we find that Meyer and Booker's cautionary advice is especially important in researching political phenomenon in developing countries.

Application: Trust in the Nicaraguan Judicial System

In this section we demonstrate that the elicited prior is useful in Bayesian political science work by applying the methodology to an analysis of citizens' trust in the judicial system using expert elicitations we personally collected in Nicaragua. We review the process by which the elicitors were selected and trained, the construction of the prior distributions, and the way that the elicited priors enter the model specification (keeping in mind Meyer and Booker's admonitions from above). Empirical work on comparative democratic institutionalism has been strictly non-Bayesian and therefore cannot systematically incorporate qualitative prior information in the way that we recommend.

The democratic consolidation literature has thus far failed to adequately address linkages between the rule of law, the judiciary, and trust in civil society. Linz and Stepan (1996), for example, emphasize the importance of a legal culture with strong roots in civil society and an impartial state apparatus, but they fail

to fully explore the dynamics of the relationship. Furthermore, not enough empirical research has been done to assess congruence between public perception and differences in elite opinion of public perception in the democratizing setting. Also, the bulk of the work in comparative politics has been in the area of general congruence between legislators and their constituencies (e.g., Pierce and Rochon 1991). The inability to establish a link between citizens' responses and elite perception of these responses outside of legislatures has been an ongoing research problem in studies of the developing world in particular.

Judges and partisan elites within the Nicaraguan political system appear to recognize that public trust in the justice system contributes to the development of a stable democratic society. For example concerning the problem of societal corruption, Filiberto Toruño, District Judge of León, asserts that "it is the challenge the judicial power must meet to gain the trust of the citizens." Because that society has traveled from authoritarianism to liberal democracy by way of a populist revolution, highly diverse ideological groups continue to exist within this society that is seeking to consolidate its democracy.⁵ Under these conditions, it is reasonable to conjecture that community elites who are aligned with the political right, community elites who are aligned with the political left, and the judicial elites will differ in their expectations concerning citizen opinion of the fairness of the justice system. *We can therefore learn about such differences in expectations of trust by contrasting models that draw elicited priors from each group of elites.* Since the priors from each group are conditioned on the same (survey) data, then posterior differences are entirely attributable to different views between the elites about what the public will say. It is important to note that we are not attempting to measure the extent to which the opinions of these various elites actually *cause* changes in public opinion (although they may), we are instead estimating differences in how these groups *view* explanations of public trust in the justice system. Thus this example highlights a substantive value of using elicited priors because these elite views are tempered by actual data on public opinion through the mechanics of Bayesian inference. Otherwise we could only descriptively report such opinion since the number of available experts is insufficient for standard statistical analysis.

The Data and Likelihood

In 2002, we administered a general survey in person to individuals in 20 community-based organizations across two cities in Nicaragua, obtaining 226 responses: 122 from León and 104 from Granada.⁶ We were purposive in picking

⁵ See for example Zatz and McDonald (1993), who argue that structural considerations constrained the Sandinistas' ability to implement a true populist revolution during the 1980s, and Williams (1994) who argues that ideological contradictions affect the Nicaraguan democratic transition.

⁶ These organizations included political parties *Partido Liberal Constitucionalista* and *FSLN*, three groups of the national women's network *AMNLAE*, the labor group *Movimiento Comunal*, the business group *Camera de Comercio*, the agricultural workers union *ATC*, evangelical Baptist churches

such organizations since their members are more likely to be knowledgeable about the judicial system and more likely to be forthcoming with their political attitudes. However, within these clusters members were selected at random to participate. This varied data acquisition approach speaks directly to concerns about the level and type of civil society activity that may exist among respondents. Such nonstochastic, multistage sampling decisions, and their associated statistical challenges, are pervasive in comparative politics (Western and Jackman 1994).

The outcome variable is a dichotomous measure of trust in justice system fairness from the public survey, with the affirmative answer coded as one. We therefore stipulate a conventional logit likelihood function for $L(\beta|\mathbf{X})$ with a random effects term (τ), but elicit expert-based priors for $p(\beta)$. The random effects component is an assumed zero-mean vector with values for each respondent included to accommodate community group heterogeneity (see Crouchley 1995 or Beck and Katz 2001 for technical details.).

Bayesian posteriors are produced by conditioning the prior on the observed data according to Bayes Law, $\pi(\beta|\mathbf{X}) \propto p(\beta)L(\beta|\mathbf{X})$, for the coefficient vector β and the data \mathbf{X} . Thus the unnormalized joint *posterior* (sampling) distribution of the parameters of interest is proportional to the prior distribution times the likelihood function, and in this way the posterior distribution combines information collected from both groups of interest.

The explanatory variables include measures of: political attitudes, ideology, occupation, and religion.⁷ The variable *Politicized* is a dichotomous measure of whether respondents view the judiciary as an inherently political institution (1 indicates “yes”). Politicization of institutions generally carries a negative connotation in Nicaraguan society, with the assumption that powerful elites benefit. To measure political ideology, respondents were categorized into three broad categories: left, center, and right. We stipulate a treatment contrast (i.e., standard “dummy” coding) with right as the ideological baseline. For occupation, we classify respondents as either blue collar, white collar, or a student. We again use a treatment contrast with student as the baseline category. Finally, there is a dichotomous explanatory variable for religion where a one indicates that the respondent is Catholic.

The Elicited Prior

Prior elicitation was performed (also in 2002) in the three described phases: deterministic, probabilistic, and informational. The first task of the deterministic

Primera Iglesia Bautista and *Iglesia Bautista Emanuel*, a teacher’s union *CGTEN* (Granada only), an activist student union *CUUN* (León only), a student business organization (León only), and a private university student organization (Granada only).

⁷Homeownership, city, gender, age, and educational level were explanatory variables that were found not to be useful in the final model. Also, interaction effects were hypothesized but not supported by the data and model.

phase was to establish the form of probability elicitation and create questions. A design matrix was built from varying the factors in the explanatory variables to produce twelve “cases” of interest from the sixty-four possible, $\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_{12}$, as suggested by Kadane et al. (1980), and described previously in the section on the normal-linear model. To create the full elicitation, we adopted a modified P-method approach that asked for percentage responses on the trust variable to these chosen levels which were then used as probability outcomes to produce point estimates with a standard logit link function as if the design matrix constituted real data. This P-method approach using hypothetical cases makes sense here because we did not want to burden our substantive experts with understanding logit coefficients. Also during this phase a written summary of the overall project was assembled to share with assessors, including: the objectives of the study, the findings of previous (1997) research, and a questionnaire with twelve sections corresponding to the twelve design points.

The probabilistic phase included three tasks: an initial interview to ascertain the assessor’s interest and knowledge concerning the project, the presentation and the assessor’s acceptance of the written elicitation survey, the collection of this survey, and the scheduling of a follow-up visit to clarify any vague information and adjust the variances. After the initial interview, eleven⁸ community elites consented to participate in the investigation: two judicial elites, three leaders of groups associated with the political right, and six leaders of groups associated with the political left.⁹

The resulting logit model standard errors from this process represent intracoder discrepancy rather than actual assessor-judged variability, so we did not yet have a full distribution for the priors. We therefore used 95% posterior credible intervals around the coefficients from the earlier (1997) fieldwork to provide initial ranges for the design-point elicitations produced here. This previous data set ($n = 73$) is a good primer for the current elicitation for three principal reasons: it was sufficiently small to produce wide credible intervals (and therefore a conservative picture of variance), it was vague enough to produce confidence in the assessor that he or she possesses better information concerning public trust,

⁸The number of experts for elicitations is necessarily a tradeoff between time and cost on one hand and the objective of averaging out any strong biases on the other. In our case, the number of available subjects who are well-informed in the subject matter *and* willing to participate was relatively small. Genest and Zidek (1986) review these issues and provide a helpful annotated bibliography to other references.

⁹The two judicial elicitations are from Maria Fabiola Betancourt, Municipal Court Judge (Granada) and Filiberto Toruño, District Court Judge (León). The three right of center civil society elicitations are from Juan Cordero, Cámara de Comercio (León); Jose Cuadra Vega, Cámara de Comercio (Granada), and Doris Juerez, Partido Liberal Constitucionalista (León). The six left of center elicitations are from Consuelo Portillo, Casa de La Mujer (Granada); Alejandro Gómez, FSLN (León); Xavier Caldera, Movimiento Comunal (Granada); Consuelo Guevarria, Movimiento Comunal (León); José Adán Zúñiga, ATC (Granada); and Nydia Loredo Perez, AMNLAE (León). The nonjudicial elites are all widely considered to be important senior leaders who are activists and organizers within these groups.

and the format of the 1997 data was a near-exact replication of the problem that assessors are asked to consider in 2002. Using these intervals around each assessor's elicitation, we allowed them to adjust these in order to reflect personal uncertainty.

All of the assessors were judged to be familiar with the idea of a 95% credible interval and its implications, so we were comfortable specifying a normal prior distribution for each group centered at the logit coefficients elicited from that group and variance equal to the average pooled elicited credible interval across groups. A *clarification meeting* constituted the informational phase of the elicitation process (it also allowed us to check for motivational or cognitive biases). While a few of the assessors adjusted estimates, the majority were strongly committed to their initial assessments. These summaries appear in Table 1 along with the posterior means from the 1997 model.

Estimation and Posterior Summary

To obtain the marginal posterior distributions, we used Gibbs sampling as implemented in the convenient package WinBUGS. We ran three separate Markov chains with overdispersed starting points for 50,000 iterations after 10,000 burn-in iterations for each of four models: three using elicited priors based on the three groups of elicitations (judicial, left, and right), and one that uses uniform priors to contrast our results with one type of a baseline. Combining the three post-burn-in independent Markov chains provides a sample size of 150,000 for *each* of the four priors, and we summarize the marginal posterior distributions in Table 2 with posterior means and 90% credible intervals (credible intervals bounded away from zero correspond to statistical significance in the conventional sense). There is strong evidence of convergence to the stationary distribution for all chains (we used many of the standard empirical diagnostic tests including: Brooks/Gelman/Rubin, Geweke, Heidelberger & Welsh, as well as graphical/visual diagnostics; See Cowles and Carlin 1996 for a detailed summary of these procedures).

TABLE 1
Summary of Elicited Priors for the Trust Model

| Respondent Characteristic | Posterior Mean 1997 | Prior Means | | | Pooled Standard Error |
|------------------------------|---------------------------|-------------------|-------------------|--------------------|-----------------------------|
| | | Judicial Group | Left of Center | Right of Center | |
| Politicized | -.9956 | -1.6090 | -1.3863 | -.9163 | (.6682) |
| Center | -.5667 | -1.2040 | -.7280 | -1.2348 | (.8722) |
| Left | -.7853 | .1500 | -.8829 | -.1919 | (.7202) |
| Blue Collar | .3825 | .3567 | .2460 | .3130 | (1.0720) |
| White Collar | .2697 | -.0513 | -.2961 | .2140 | (.8246) |
| Catholic | -1.1748 | -1.6090 | .5142 | -.2730 | (.7441) |

TABLE 2
Nicaragua Posterior Results: Elicited Categorical Groupings

| Respondent Characteristic | Judicial Prior | | Left Prior | | Right Prior | | Uniform Prior | |
|------------------------------|----------------|---------------|------------|--------------|-------------|--------------|---------------|--------------|
| | Coef. | 90% CI | Coef. | 90% CI | Coef. | 90% CI | Coef. | 90% CI |
| (Intercept) | .752 | [.09:1.41] | .582 | [−.05:1.21] | .492 | [−.15:1.14] | .623 | [−.27:1.51] |
| Politicized | −1.573 | [−2.11:−1.03] | −1.397 | [−1.90:−.90] | −1.408 | [−1.92:−.90] | −1.638 | [−2.38:−.90] |
| Center | −.554 | [−1.13:.02] | −.569 | [−1.13:−.01] | −.597 | [−1.17:−.02] | −.573 | [−1.29:.14] |
| Left | −.740 | [−1.32:−.16] | −.859 | [−1.42:−.30] | −.823 | [−1.39:−.25] | −.983 | [−1.76:−.20] |
| Blue Collar | .912 | [.32:1.50] | .768 | [.20:1.34] | 1.053 | [.47:1.64] | 1.143 | [.36:1.93] |
| White Collar | .590 | [.02:1.16] | .439 | [−.12:1.00] | .702 | [.13:1.27] | .783 | [.06:1.51] |
| Catholic | −.732 | [−1.26:−.20] | −.354 | [−.85:.15] | −.475 | [−1.00:.03] | −.577 | [−1.24:.09] |
| <i>Var</i> (τ) | 4.624 | [3.90:5.35] | 4.630 | [3.90:5.36] | 4.641 | [3.92:5.37] | 4.604 | [3.88:5.33] |

All three prior elicitations lead to a reliable conclusion that the extent to which the judiciary is perceived as an inherently political institution (Politicized), it erodes trust in justice system fairness. Also all three priors imply that ideologically left citizens have greater mistrust in the fairness of the justice system as compared to ideologically right citizens, and there is the common implication that blue collar workers have greater trust in justice system fairness than both students and white collar workers. While our assessors are representative of different aspects of Nicaraguan society, the similarity of these findings indicates that the groups have closely aligned views about these three explanations of public trust in the justice system.

Although all three marginal posteriors for Politicized are negative and reliably distant from zero, it is interesting that the one produced from the judicial prior is the largest in magnitude. Possibly this reflects a slight preference for societal consensus over ideological polarization in government control. That is, judicial officials appear to believe that this factor deteriorates trust more than do community organization leaders of either orientation (their posterior means are close enough to each other that a left-right distinction seems unwarranted).

The occupational variables all produce positive posterior means across the three elicited prior forms, with only one (White Collar for the left prior) that fails to be statistically defensible. Since the student group is the reference category, this means that there is general agreement across the elicitation groups that those traditionally employed are more trusting than students. In the two cases where we can make comparisons, judicial prior and right prior, the posterior means are higher for blue collar workers than for white collar workers. Although the differences are roughly equivalent, the right prior has a slightly higher effect here suggesting that this group focuses more on employment status than the judicial elicitors.

There are some more pronounced differences in posterior findings across elicitations. Both the judicial prior and the right prior lead to a positive and statistically supportable relationship between white collar occupational status and trust in justice system fairness. This finding suggests these groups hold the view that citizens with (typically) higher income and status have higher trust. Yet when using the elicited prior from the ideological left, there is no evidence to support such a view from the model. This may indicate possible cynicism about the relationship (or perhaps just some level of underestimation) from the left assessors and could reflect their personal dissatisfaction with the center-right Partido Liberal Constitucionalista currently controlling government.

The left prior and right prior produce coefficient 90% credible intervals bounded away from zero for the (negative) relationship between ideologically centrist respondents and trust in the justice system compared to ideologically right respondents (the baseline group). The judicial prior, however, produces no supportable evidence for claiming such a relationship. Interestingly, the judicial prior does support a distinction between left respondents and right respondents, indicating that moving leftwards in ideology should reduce trust in government

(also found with the left and right priors). Taken as a whole, this all suggests that judges see less nuance on the ideological dimension, and focus more on citizens' perceived politicization of institutions as determinants of trust. Conversely, the left and right priors imply more of a *scale* of gradually declining trust moving across three-points from right respondents to center respondents and then to left respondents.

The judicial prior model also differs from the other models in that it is the only one that produces a supportable relationship between being Catholic and disinclination to trust the justice system (relative to non-Catholics of course). Judges, who appear to think in more institutional than ideological terms, apparently consider Catholic citizens less trustful of the justice system than non-Catholics. Since Catholics are the vast majority (about 85% of the Nicaraguan population though not all are observant), this is really a statement about the judicial view of public opinion in general.

Using the posterior predictive mean¹⁰ from each model offers evidence that ideology has a similar effect in some scenarios. For instance, an ideologically central, Catholic student who believes that the judicial system is politicized has essentially the same expected probability of trust across all four priors (including the uniform discussed below), $p = .12$, which indicates wide agreement from the experts for this particular case (this similarity is partly a function of the magnitude of the Politicized posterior mean across all four models).

Those still skeptical about informed priors will appreciate the inclusion of a model with uniform priors in Table 2 as a way to express the implications of prior ignorance. Recall that this distributional form is only one of several ways that Bayesian models express prior uninformedness. The first thing to notice is that the marginal posterior summaries from this very different prior are generally similar to the strongly informed varieties. That is, none of the posterior means from the uniform prior are even close to an order of magnitude away from the others, no signs are switched, and the posterior variances are necessarily larger but not terrifically so. This informal sensitivity analysis suggests a high level of robustness with regard to prior specification for the elicited approach. We also employed a tool of Bayesian robustness by "contaminating" (mixing) our priors with very diffuse forms as suggested by Berger (1985, section 4.7.4). In doing so we saw little sensitivity even to moderate proportions of contamination priors on each of our elicited forms, and they still differed noticeably.

Some minor differences, however, can be found in the size of the coefficients from applying the uniform prior. For the three variables where all four priors lead to marginal posteriors with 90% credible intervals bounded away from zero (Politicized, Left, Blue Collar), the uniform prior posteriors have slightly higher

¹⁰ This is the mean of the posterior predictive distribution for a hypothetical new data point, x_{new} , after the data, \mathbf{X} , have been observed: $p(x_{new}|\mathbf{X}) = \int_{\beta} p(x_{new}|\beta|\mathbf{X})d\beta = \int_{\beta} p(x_{new}|\beta)p(\beta|\mathbf{X})d\beta$. So this distribution is the product of the single variable PDF or PMF times the full data likelihood where we integrate over uncertainty in β to give a probability statement that is dependent on the observed data only.

posterior means (in absolute value terms). Thus, although the groups of assessors come from differing perspectives, the elicited priors produce a dampening effect on the influence of the explanatory variables relative to a low-information prior model. Therefore in the absence of substantive prior knowledge, the model would exaggerate the importance of these factors, and we would be inclined to overstate our results in the subsequent discussion.

Conclusion: Elicited Priors in Political Science Research

In our Nicaragua research we sought to combine elite opinion with standard survey data in a single model specification. The goal was to contrast elite views of the relationship between important covariates and trust in the judiciary. So this is a good example of how elicited priors can be useful since it allows us to measure and incorporate qualitative data from differing, but comparable, sources into the same statistical specification. Obviously there are informal and ad hoc means of doing this, but researchers who wish to systematically combine qualitative and quantitative information in the same model have found few helpful procedures thus far. To address this problem we have explained four general methodologies to format elicited information into a standard empirical model. Such tools are relatively new in Bayesian statistics and entirely new in political science.

We do not want to leave the impression that prior elicitation is a panacea for the social sciences. In addition to requiring a full Bayesian setup with justified informative priors, there has to be a substantive motivation for including specific information outside of the sample. If this is not an important consideration, then more traditional approaches to prior specification are preferred. Furthermore, elicitation is generally produced from detailed fieldwork, although this requirement plays into a traditional strength of qualitative scholars in political science. Elicited priors can also be generated from historical document analysis, but this is often considerably more difficult and the topic is beyond our scope here (see Spiegelhalter, Abrams, and Myles [2004], Chapter 5).¹¹

Substantively focused priors can be particularly helpful in comparative politics research where data are often small samples picked nonrandomly by researchers because they contain deep substantive information. Western and Jackman (1994) observe that the associated problem with using classic statistical approaches here is that such prior information is “useful in the comparative context where rich historical material is commonly available, sparking ideas for researchers, but formally discarded in the final analysis” (1994, 412). Unfortunately the emphasis in subsequent Bayesian work in political science has been

¹¹ The real distinction is that with historical document analysis subjects cannot be interactively queried, and therefore this is technically not “eliciting” in the sense that we have described. A famous example of drawing priors from document analysis is Mosteller and Wallace’s (1963) analysis of the twelve Federalist Papers with unknown authorship where they construct weighted prior distributions of word usage for both Madison and Hamilton from Federalist Papers known to have been written by each.

on *noninformed* priors and is thus no more helpful for pulling together qualitative and quantitative information than frequentist statistical approaches that ignore prior information altogether (although this approach can address intractable frequentist models and it does move summary analysis into probabilistic terms). Berger also warns that “noninformative prior Bayesian analyses are not automatically sensible” (1985, 231) in the sense of portraying ignorance since they are not *guaranteed* to minimally affect the posterior. Elicited priors, and priors derived from substantive knowledge in general, provide the opposite effect by tapping into the extensive background information that political scientists bring to empirical research projects. Therefore an undesired tradeoff between generalizability and rich detail is avoided and researchers can enjoy the benefits of both perspectives. Currently there is no other methodology in political science that proposes to accomplish this integration.

Acknowledgment

We thank Janet Box-Steffensmeier, Jack Buckley, Suzie De Boef, James Fowler, John Freeman, Bob Jackman, Andrew Martin, Lee Sigelman, Rick Waterman, and Chris Zorn for helpful comments.

Manuscript submitted June 8, 2004

Final manuscript received March 2, 2005

References

- Al-Awadhi, S. A., and P. H. Garthwaite. 2001. “Prior Distribution Assessment for a Multivariate Normal Distribution: An Experimental Study.” *Journal of Applied Statistics* 28(1): 5–23.
- Ascher, W. 1978. *Forecasting: An Appraisal for Policymakers and Planners*. Baltimore: Johns Hopkins University Press.
- Ayyub, Bilal M. 2001. *Elicitation of Expert Opinions for Uncertainty and Risks*. New York: CRC Press.
- Bartels, Larry. 1996. “Pooling Disparate Observations.” *American Journal of Political Science* 40(3): 905–42.
- Beck, Nathaniel, and Simon Jackman. 1998. “Beyond Linearity by Default: Generalized Additive Models.” *American Journal of Political Science* 42(2): 596–627.
- Beck, Nathaniel, and Jonathan N. Katz. 2001. “Random Coefficient Models for Time-Series–Cross-Section Data: The 2001 Version.” Available at: <http://polmeth.wustl.edu/>.
- Bedrick, Edward J., Ronald Christensen, and Wesley O. Johnson. 1997. “Bayesian Binomial Regression: Predicting Survival at a Trauma Center.” *The American Statistician* 51(3): 211–18.
- Bennett, S. E. 2002. “‘Perestroika’ Lost: Why the Latest ‘Reform’ Movement in Political Science Should Fail.” *PS: Political Science & Politics* 35(2): 177–79.
- Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. New York: Springer-Verlag.
- Berk, Richard A., Bruce Western, and Robert E. Weiss. 1995. “Statistical Inference for Apparent Populations.” *Sociological Methodology* 25: 421–58.
- Blockley, D. I. 1979. “The Calculations of Uncertainty in Civil Engineering.” *Proceedings of the Institute of Civil Engineering* 67 (Part 2): 313–26.

- Brady, Henry E., and David Collier. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.
- Bunn, D. W. 1978. "The Estimation of a Dirichlet Prior Density." *Omega* 6: 371–73.
- Bunn, D. W. 1979. "Estimation of Subjective Probability Distributions in Forecasting and Decision Making." *Technical Forecasting and Social Change* 14: 205–16.
- Carlin, Bradley P., Kathryn Chaloner, Timothy Church, Thomas A. Louis, and John P. Matts. 1993. "Bayesian Approaches for Monitoring Clinical Trials with an Application to Toxoplastic Encephalitis Prophylaxis." *The Statistician* 42(4): 355–67.
- Carlin, Bradley P., Kathryn Chaloner, T. A. Louis, and F. S. Rhames. 1995. "Elicitation, Monitoring, and Analysis for an AIDS Clinical Trial." In *Case Studies in Bayesian Statistics, Volume II*, eds. Constantine Gatsonis, James S. Hodges, Robert E. Kass, and Nozer D. Singpurwalla. New York: Springer-Verlag, pp. 48–78.
- Chaloner, Kathryn M., Timothy Church, Thomas A. Louis, and John P. Matts. 1993. "Graphical Elicitation of a Prior Distribution for a Clinical Trial." *The Statistician* 42(4): 341–53.
- Chaloner, Kathryn M., and G. T. Duncan. 1983. "Assessment of a Beta Prior Distribution: PM Elicitation." *The Statistician* 32(1/2): 174–80.
- Chaloner, Kathryn M., and G. T. Duncan. 1987. "Some Properties of the Dirichlet-Multinomial Distribution and Its Use in Prior Elicitation." *Communications in Statistics—Theory and Methods* 15: 511–23.
- Cooke, Roger M. 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York: Oxford University Press.
- Cosmides, L., and J. Tooby. 1996. "Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions from the Literature on Judgment Under Uncertainty." *Cognition* 58(1): 1–73.
- Cowles, M. K., and B. P. Carlin. 1996. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review." *Journal of the American Statistical Association* 91(434): 883–904.
- Crosby, Michael A. 1980. "Implications of Prior Probability Elicitation on Auditor Sample-Size Decisions." *Journal of Accounting Research* 18(2): 585–93.
- Crouchley, Robert. 1995. "A Random-Effects Model for Ordered Categorical Data." *Journal of the American Statistical Association* 90(430): 489–98.
- Dong, W-M., and F. S. Wong. 1986a. "From Uncertainty to Approximate Reasoning: Part 1: Conceptual Models and Engineering Interpretations." *Civil Engineering Systems* 3: 143–54.
- Dong, W-M., and F. S. Wong. 1986b. "From Uncertainty to Approximate Reasoning: Part 2: Reasoning with Algorithmic Rules." *Civil Engineering Systems* 4: 192–202.
- Dong, W-M., and F. S. Wong. 1986c. "From Uncertainty to Approximate Reasoning: Part 3: Reasoning with Conditional Rules." *Civil Engineering Systems* 5: 112–21.
- Fisher, Ronald A. 1922. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Statistical Society A* 222: 309–68.
- Francisco, Ronald A. 1996. "Coercion and Protest: An Empirical Test in Two Democratic States." *American Journal of Political Science* 40(4): 1179–1204.
- Freedman, L. S., and D. J. Spiegelhalter. 1983. "The Assessment of Subjective Opinion and Its Use in Relation to Stopping Rules for Clinical Trials." *The Statistician* 32(1/2): 153–60.
- Furuta, H., K. S. Fu, and J. T. P. Yao. 1985. "Structural Engineering Applications of Expert Systems." *Computer Aided Design* 17(9): 410–19.
- Garthwaite, Paul H., and James M. Dickey. 1988. "Quantifying Expert Opinion in Linear Regression Problems." *Journal of the Royal Statistical Society, Series B* 50(3): 462–74.
- Garthwaite, Paul H., and James M. Dickey. 1992. "Elicitation of Prior Distributions for Variable Selection Problems in Regression." *Annals of Statistics* 20(4): 1697–1719.
- Gavaskar, Umesh. 1988. "A Comparison of Two Elicitation Methods For a Prior Distribution For a Binomial Parameter." *Management Science* 34(6): 784–90.
- Genest, Christian, and James V. Zidek. 1986. "Combining Probability Distributions: A Critique and an Annotated Bibliography." *Statistical Science* 1(1): 114–35.

- Gerber, Elisabeth R., and John E. Jackson. 1993. "Endogenous Preferences and the Study of Institutions." *American Political Science Review* 87(3): 639–56.
- Gill, Jeff. 2002. *Bayesian Methods for the Social and Behavioral Sciences*. New York: Chapman & Hall.
- Groves, Robert M. 1987. "Research on Survey Data Quality." *Public Opinion Quarterly* 51(2): 156–72.
- Helmer, O. 1967. *Analysis of the Future: The Delphi Method*. Santa Monica, CA: Rand Corporation.
- Hill, J., and Kriesi, H. 2001. "Classification by Opinion Changing Behavior: A Mixture Model Approach." *Political Analysis* 4: 301–24.
- Hogarth, Robin M. 1975. "Cognitive Processes and the Assessment of Subjective Probability Distributions." *Journal of the American Statistical Association* 70(350): 271–89.
- Hora, S. C., J. A. Hora, and N. G. Dodd. 1992. "Assessment of Probability Distributions for Continuous Random Variables: A Comparison of the Bisection and Fixed-value Methods." *Organizational Behavior and Human Decision Processes* 51(1): 133–55.
- Ibrahim, J. G., and Chen, M-H. 2000. "Prior Elicitation and Variable Selection for Generalized Linear Mixed Models." In *Generalized Linear Models: A Bayesian Perspective*, eds. Dipak K. Dey, Sujit K. Ghosh, and Bani K. Mallick. New York: Marcel Dekker, pp. 41–53.
- Jackman, Simon. 2000a. "Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation." *Political Analysis* 8(4): 307–32.
- Jackman, Simon. 2000b. "Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo." *American Journal of Political Science* 44(2): 375–404.
- Jackman, Simon. 2001. "Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking." *Political Analysis* 9(3): 227–41.
- Jeffreys, Harold. 1961. *The Theory of Probability*. Oxford: Oxford University Press.
- Johnson, Valen E., and James H. Albert. 1999. *Ordinal Data Modeling*. New York: Springer-Verlag.
- Judd, Charles M., and Gary H. McClelland. 1984. "Measurement." In *The Handbook of Social Psychology*, eds. Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey. New York: McGraw Hill, pp. 180–232.
- Kadane, Joseph B. 1980. "Predictive and Structural Methods for Eliciting Prior Distributions." In *Bayesian Analysis in Econometrics and Statistics*, ed. A. Zellner. Amsterdam: North Holland, pp. 89–93.
- Kadane, J. B. 1986. "Progress Toward a More Ethical Method Clinical Trials." *Journal of Medical Philosophy* 11: 385–405.
- Kadane, Joseph B., James M. Dickey, Robert L. Winkler, Wayne S. Smith, and Stephen C. Peters. 1980. "Interactive Elicitation of Opinion for a Normal Linear Model." *Journal of the American Statistical Association* 75(372): 845–54.
- Kadane, Joseph B., and Robert L. Winkler. 1988. "Separating Probability Elicitation from Utilities." *Journal of the American Statistical Association* 83(402): 357–63.
- Kadane, Joseph B., and Lara J. Wolfson. 1998. "Experiences in Elicitation." *Journal of The Royal Statistical Society, Series D* 47(1): 3–19.
- Kahneman, Daniel, and Amos Tversky. 1982. "Subjective Probability: A Judgement of Representativeness." In *Judgment Under Uncertainty: Heuristics and Biases*, eds. D. Kahneman, P. Slovic, and A. Tversky. New York: Cambridge University Press, pp. 43–47.
- Kass, Robert E., and Joel B. Greenhouse. 1989. "Comments on the paper by J. H. Ware." *Statistical Science* 4(4): 310–17.
- Kasza, Gregory J. 2000. "'Technicism' Supplanting Disciplinarity among Political Scientists." *PS: Political Science & Politics* 33(4): 737–38.
- King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.
- Leamer, Edward E. 1972. "A Class of Informative Priors and Distributed Lag Analysis." *Econometrica* 40(6): 1059–81.

- Leamer, Edward E. 1992. "Bayesian Elicitation Diagnostics." *Econometrica* 60(4): 919–42.
- Lindley, Dennis V. 1961. "The Use of Prior Probability Distributions in Statistical Inference and Decision." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, pp. 453–68.
- Lindley, Dennis V., and Nozer D. Singpurwalla. 1991. "On the Evidence Needed to Reach Agreed Action Between Adversaries, With Application to Acceptance Sampling." *Journal of the American Statistical Association* 86(416): 933–97.
- Lindley, D. V., and A. F. M. Smith. 1972. "Bayes Estimates for the Linear Model." *Journal of the Royal Statistical Society, Series B* 34(1): 1–41.
- Lindley, D. V., A. Tversky, and R. V. Brown. 1979. "On the Reconciliation of Probability Assessments." *Journal of the Royal Statistical Society, Series A* 142(2): 146–80.
- Lin, J., and A. Stepan. 1996. *Problems of Democratic Transition and Consolidation*. Baltimore: Johns Hopkins University Press.
- Martin, Andrew D., and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10(2): 134–53.
- Meyer, Mary A., and Jane M. Booker. 2001. *Eliciting and Analyzing Expert Judgement: A Practical Guide*. Philadelphia: SIAM Press.
- Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association* 58(302): 275–309.
- Neyman, Jerzy. 1937. "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." In *A Selection of Early Statistical Papers of J. Neyman*. Berkeley: University of California Press, pp. 333–80.
- O'Hagan, A. 1998. "Eliciting Expert Beliefs in Substantial Practical Applications." *The Statistician* 47(1): 21–35.
- Pearson, Karl. 1920a. "The Fundamental Problem of Practical Statistics." *Biometrika* 13(1): 1–16.
- Pearson, Karl. 1920b. "Note on 'The Fundamental Problem of Practical Statistics.'" *Biometrika* 13: 300–301.
- Pierce, Roy, and Thomas R. Rochon. 1991. "Constancy of Legislative Perceptions of Constituency Opinion: French Socialist Candidates in 1967 and 1978." *Comparative Political Studies* 23(3): 478–96.
- Press, S. James. 1989. *Bayesian Statistics: Principles, Models, and Applications*. New York: John Wiley & Sons.
- Quinn, Kevin M., Andrew Martin, and Andrew B. Whitford. 1999. "Voter Choice in Multi-Party Democracies: A Test of Competing Theories and Models." *American Journal of Political Science* 43(4): 1231–47.
- Raiffa, Howard, and Robert Schlaifer. 1961. *Applied Statistical Decision Theory*. Cambridge: Harvard School of Business Administration.
- Savage, Leonard J. 1971. "Elicitation of Personal Probabilities and Expectations." *Journal of the American Statistical Association* 66(336): 783–801.
- Savage, Leonard J. 1972. *The Foundations of Statistics*. New York: Dover Publications.
- Seidenfeld, Teddy. 1985. "Calibration, Coherence, and Scoring Rules." *Philosophy of Science* 52: 274–94.
- Smith, Alastair. 1999. "Testing Theories of Strategic Choice: The Example of Crisis Escalation." *American Journal of Political Science* 43(4): 1254–83.
- Smith, W., F. Finn, C. Saraf, and R. Kulkarni. 1978. "Bayesian Analysis Methodology for Verifying Recommendations to Minimize Asphalt Pavement Distress." Final Report 9–4A. National Cooperative Highway Research Program.
- Spetzler, Carl S., and Carl-Axel S. Staël von Holstein. 1975. "Probability Encoding in Decision Analysis." *Management Science* 22: 340–58.
- Spiegelhalter, David J., Keith R. Abrams, and Jonathan P. Myles. 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: John Wiley & Sons.

- Spiegelhalter, David J., Laurence S. Freedman, and Mahesh K. B. Parmar. 1994. "Bayesian Approaches to Randomized Trials." *Journal of the Royal Statistical Society, Series A* 157(3): 357–416.
- Steffey, Duane. 1992. "Hierarchical Bayesian Modeling with Elicited Prior Information." *Communications in Statistics—Theory and Methods* 21(3): 799–821.
- Tiao, G. C., and A. Zellner. 1964. "Bayes' Theorem and the Use of Prior Knowledge in Regression Analysis." *Biometrika* 51(1/2): 219–30.
- Tversky, Amos, and Daniel Kahneman. 1983. "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement." *Psychological Review* 90(4): 293–315.
- Western, Bruce. 1998. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach." *American Journal of Political Science* 42(4): 1233–59.
- Western, Bruce, and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88(2): 412–23.
- Williams, Phillip J. 1994. "Dual Transitions From Authoritarian Rule: Popular and Electoral Democracy in Nicaragua." *Comparative Politics* 26(2): 169–86.
- Winkler, Robert L. 1967. "The Assessment of Prior Distributions in Bayesian Analysis." *Journal of the American Statistical Association* 62(319): 776–800.
- Zatz, M. S., and J. H. McDonald. 1993. "Structural Contradictions and Ideological Consistency: Changes in the Form and Content of Cuban Criminal Law." In *Making Law: The State, The Law, and Structural Contradictions*, eds. William Chamblis and Marjorie Zatz. Bloomington: Indiana University Press, pp. 127–67.

Jeff Gill (jgill@ucdavis.edu) is associate professor of political science, University of California–Davis, Davis, CA 95616.

Lee D. Walker (lee.walker@uky.edu) is assistant professor of political science, University of Kentucky, Lexington, KY 40506.