*Specification Uncertainty and Model Averaging**

Larry M. Bartels, *Princeton University*

Theory: Data analysts sometimes report (and more often produce) results from many alternative models with different explanatory variables, functional forms, observations, or exogeneity assumptions. Classical statistical theory is ill-suited to make sense of this practice.
Hypotheses: Bayesian statisticians have recently proposed a coherent procedure for taking account of specification uncertainty by averaging results from a variety of different model specifications. The model-averaging procedure has the general effect of discounting evidence derived from elaborate specification searches, especially when alternative models produce markedly different results.
Methods: I describe the model-averaging procedure, and illustrate its application using examples drawn from a controversy in comparative political economy between Lange and Garrett (1985, 1987) and Jackman (1987), and from the work of Erikson, Wright, and McIver (1993) on public opinion and policy in the American states. In addition, I propose two classes of reference priors that might usefully supplement the uniform model priors typically adopted in model averaging—a ''dummy-resistant prior'' for dealing with outlier observations, and a family of ''search-resistant priors'' for representing sequential specification searches.
Results: The model-averaging procedure seems to offer a convenient approximation to full-blown Bayesian analysis in typical social science settings. It is simple to implement, and uses the variety of alternative model specifications already being produced by data analysts to shed some useful light on the inferential implications of specification uncertainty.

There is an embarrassing gulf between the elegant superstructure of statistical theory and the actual practices of real data analysts. As Achen (1982, 16) put it,

*This article is a long addendum to my short discussion of ''Five Approaches to Model Specification'' in *The Political Methodologist* (Bartels 1990). I am grateful to Christopher Achen and Nathaniel Beck for relevant inspiration, to Simon Jackman and John Londregan for relevant conversation, and especially to Bruce Western for sharing and discussing his work on ''Model Uncertainty in Macrosociology'' (Western 1996), which has greatly influenced my own work on this topic. Earlier versions of this article were presented at the Twelfth Annual Political Methodology Summer Meeting at Indiana University, July 1995, and at the Annual Meeting of the American Political Science Association, Chicago, September 1995. Christopher Achen, Henry Brady, Robert Erikson, John Freeman, Gary King, Walter Mebane, and John Zaller provided especially helpful comments and criticism.

American Journal of Political Science, Vol. 41, No. 2, April 1997, Pp. 641–674

> conventional statistical methods require strong and precise assumptions about the functional relationship among the variables and the behavior of unmeasured causes. In social science applications, these postulates are not supplied by theory. The ensuing logical gap is the principal obstacle to social data analysis and the most challenging intellectual problem facing the social science methodologist.

In the absence of the "strong and precise assumptions" required by statistical theory, data analysts typically engage in complex "specification searches" (Leamer 1978) employing a variety of alternative—mutually contradictory—strong and precise assumptions. They may run a plausible regression, modify their model on the basis of the regression results, run another regression, and so on, almost (but not quite) *ad infinitum.* Modifications may include dropping variables whose coefficients are "insignificant" or have the "wrong" (unexpected) sign, deleting observations that seem to deviate markedly from the main patterns in the data, replacing apparently poor measures with potentially better measures of the same theoretical constructs, or adding more elaborate sets of explanatory variables if the data seem cooperative enough to confess further secrets.

It is obviously foolish to treat the resulting parameter estimates as though they were produced in conformity with traditional statistical theory. None of the standard results derived under the assumption that the correct statistical model is known with certainty can be justified when that assumption is relaxed. Indeed, in many instances the standard results do not even hold to a rough approximation for models resulting from more or less complex specification searches (Adams 1991; Freedman 1983; Green 1990; Judge and Bock 1978).[1] It is hard to avoid the conclusion that the whole conventional statistical armamentarium of confidence intervals and hypothesis tests is—or at least should be—almost irrelevant to real data analysis.

Wise data analysts recognize the mismatch between theory and practice, and do their best to overcome the limitations of statistical theory by interpreting the results of their analyses in complicated ways only loosely based on formal calculations of confidence intervals and hypothesis tests. They strive to learn (and convey) what their data have to say without succumbing to the familiar pitfalls of "barefoot empiricism," "data mining," and "fishing." They explore (and report) the sensitivity of their conclusions to plausible alternative model assumptions. And they recognize that, as Leamer (1978, 13) noted, "the apparent statistical evidence implied by the

[1]For example, Adams (1991) tested 114 distinct regression model selection strategies on random data with from 10 to 70 observations and from 5 to 30 predictors with intercorrelations ranging from zero to .75. All of the stepwise or goodness-of-fit based strategies produced nominal p-values below .01, whereas the true p-values with random data were .50.

final equation must be discounted; the greater the range of search, the greater must be the discount.''

So far, so good. But how, exactly, should that discount be calculated? That is the subject of this article. I describe a new technique developed by Bayesian statisticians (most notably Draper 1995 and Raftery 1995, building upon the work of Jeffreys 1961, Leamer 1978, and others) in which the results of various alternative model specifications are averaged to produce inferences that make explicit allowance for the implications of specification uncertainty. This technique has the general effect of discounting evidence derived from elaborate specification searches, especially when alternative models produce markedly different results. Thus, unlike more conventional model selection techniques which aim to identify a single ''best'' model but then treat the results of that model as if it was the only one examined, the model averaging procedure produces both parameter estimates and standard errors that honestly reflect the observed variation of results across a range of plausible models.

The model-averaging approach to specification uncertainty represents a formalization of enlightened data-analytic practice along explicit Bayesian lines.[2] In order to gauge its potential utility, I focus upon two examples of real social science data analysis—a controversy in comparative political economy between Lange and Garrett (1985, 1987) and Jackman (1987), and an analysis by Erikson, Wright, and McIver (1993) of the impact of public opinion on policy outcomes in the American states. A detailed consideration of these two examples illustrates several important strengths and limitations of the model-averaging approach. It suggests that model averaging is a marked improvement over conventional statistical practice, and an attractive (though imperfect) approximation to full-blown Bayesian analysis.

The Logic of Model Averaging

Our aim is to make an inference about some quantity of interest, Δ, on the basis of some observed data, $\mathbf{X}$ and $\mathbf{y}$. (The quantity Δ may be a regression parameter, a forecast of some future observation, or some other quantity of interest.) In classical statistical theory, the data are only informa-

[2]Western and Jackman (1994) provided an elementary introduction to Bayesian regression analysis in political science. The Bayesian approach offers a useful starting point for tackling the problem of specification uncertainty because it is designed precisely to incorporate uncertain non-sample information of the sort provided by social science theory. As will be evident here, subsequent developments have done much to justify Leamer's (1978, 2) assertion that ''the Bayesian approach is sufficiently flexible that, with suitable alterations, specification searches can be made legitimate, or at least understandable. This does not seem to be the case with the classical model of inference.''

tive about Δ when they are interpreted in light of a definite statistical model, M, which specifies a relevant population, a set of relevant variables, the functional form of the relationships among these variables, and the nature of relevant stochastic influences. All of our inferences about Δ depend not only upon the data $\mathbf{X}$ and $\mathbf{y}$, but also upon the assumptions embodied in the model M: in the language of Bayesian statistical theory, they are derived from conditional posterior distributions of the form $p(\Delta|\mathbf{X}, \mathbf{y}, M)$.

To obtain a posterior distribution that depended upon the data $\mathbf{X}$ and $\mathbf{y}$ but not upon a specific model M, we would have to treat M as a nuisance parameter and integrate it out, producing the unconditional posterior distribution

$$p(\Delta|\mathbf{X}, \mathbf{y}) = \int p(\Delta|\mathbf{X}, \mathbf{y}, M)p(M|\mathbf{X}, \mathbf{y})dM. \qquad [1]$$

This approach is infeasible both in principle and in practice, since the set of conceivable models for which we would need to compute conditional posterior distributions of the form $p(\Delta|\mathbf{X}, \mathbf{y}, M)$ is not a set of finite measure. However, we can at least reduce our dependence upon model assumptions by integrating over a discrete set of J alternative models $M_1, \ldots M_j, \ldots M_J$ representing the most plausible, salient, or otherwise interesting alternative sets of statistical assumptions. The resulting expression for the unconditional posterior distribution corresponding to Equation [1] is

$$p(\Delta|\mathbf{X}, \mathbf{y}) = \sum_j p(\Delta|\mathbf{X}, \mathbf{y}, M_j)\, p(M_j|\mathbf{X}, \mathbf{y}), \qquad [2]$$

which is a simple weighted average of the conditional posterior distributions $p(\Delta|\mathbf{X}, \mathbf{y}, M_1), \ldots p(\Delta|\mathbf{X}, \mathbf{y}, M_j), \ldots p(\Delta|\mathbf{X}, \mathbf{y}, M_J)$ for the alternative models $M_1, \ldots M_j, \ldots M_J$, each weighted by the corresponding posterior model probability $p(M_j|\mathbf{X}, \mathbf{y}) \equiv \pi_j$, with $\Sigma_j\, \pi_j = 1$.[3]

The mean and variance of the unconditional posterior distribution in Equation [2] depend upon the means and variances of the corresponding conditional posterior distributions and upon the posterior model probabilities π_j (Leamer 1978, 118). The mean of the unconditional posterior distribution is

$$E(\Delta|\mathbf{X}, \mathbf{y}) = \sum_j \pi_j\, E(\Delta|\mathbf{X}, \mathbf{y}, M_j) \qquad [3]$$

[3]Obviously, the assumption that only J discrete models get positive probability, so that $\Sigma_j\, \pi_j = 1$, is a crucial simplification. I discuss its implications in the Discussion section below.

and the variance of the unconditional posterior distribution is

$$V(\Delta|\mathbf{X}, \mathbf{y}) = \sum_j \pi_j V(\Delta|\mathbf{X}, \mathbf{y}, M_j) + \sum_j \pi_j [E(\Delta|\mathbf{X}, \mathbf{y}, M_j) - E(\Delta|\mathbf{X}, \mathbf{y})]^2. \qquad [4]$$

The mean of the unconditional posterior distribution in Equation [3] is a simple weighted average of the conditional posterior means, while the variance of the unconditional posterior distribution in Equation [4] has two components, the first a weighted average of the conditional posterior variances and the second a weighted average of the squared deviations between the conditional and unconditional posterior means.

My focus in this article is on the coefficients of a linear regression model. With a diffuse normal-gamma prior distribution over the regression parameters, the mean of the posterior distribution for a parameter $\boldsymbol{\beta}_j$ conditional upon model M_j is simply the corresponding least squares parameter estimate $\mathbf{b}_j$ (or zero if the parameter $\boldsymbol{\beta}_j$ is omitted from model M_j), and the variance of the conditional posterior distribution is simply the variance $V(\mathrm{b}_j)$ of the least squares parameter estimate (or zero if $\boldsymbol{\beta}_j$ is omitted from model M_j).[4] Thus, the mean of the unconditional posterior distribution corresponding to Equation [3] is

$$E(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) \equiv \bar{\mathbf{b}} = \sum_j \pi_j \mathbf{b}_j \qquad [5]$$

and the variance of the unconditional posterior distribution corresponding to Equation [4] is

$$V(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) = \sum_j \pi_j V(\mathbf{b}_j) + \sum_j \pi_j (\mathbf{b}_j - \bar{\mathbf{b}})^2, \qquad [6]$$

where, as before, $\pi_j \equiv p(M_j|\mathbf{X}, \mathbf{y})$ is the posterior probability for model M_j.

With diffuse priors over the parameters of the various alternative re-

[4]The normal-gamma prior is mathematically convenient (since with normally distributed disturbances the posterior distributions of $\boldsymbol{\beta}_j$ and σ_j must be in the same families as the corresponding prior distributions) and provides a rationale for treating the least squares parameter estimate $\mathbf{b}_j$ as the posterior mean of $\boldsymbol{\beta}_j$ given a diffuse prior (Leamer 1978, 78–9). Different priors might be desirable in some instances, but would complicate the analysis significantly.

gression models, the unconditional posterior mean for each coefficient is simply an average of the separate parameter estimates produced by the various models, each weighted by the posterior probability associated with that model. The unconditional posterior variance for each coefficient is the same weighted average of the separate variances of the parameter estimates in the various alternative models, plus an additional weighted average representing uncertainty resulting from the variability of the parameter estimates across the range of alternative models.[5] This additional weighted average may make the unconditional posterior variance large due to specification uncertainty, even if the various alternative models all produce very precise parameter estimates, if those estimates differ significantly from one model to another.

The only significant remaining difficulty is to calculate the posterior model probabilities $\pi_j \equiv p(M_j|\mathbf{X}, \mathbf{y})$. By Bayes' theorem,

$$\pi_j \equiv p(M_j|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{X}, \mathbf{y}|M_j)\, p(M_j), \qquad [7]$$

where $p(M_j) \equiv \pi_j^0$ is the prior probability of model M_j and

$$p(\mathbf{X}, \mathbf{y}|M_j) = \int p(\mathbf{X}, \mathbf{y}|\boldsymbol{\theta}_j, M_j)\, p(\boldsymbol{\theta}_j|M_j)\, d\boldsymbol{\theta}_j \qquad [8]$$

is the *marginal likelihood* of the data under model M_j.

When alternative models are being compared, the extent to which the data favor model M_i over model M_j is indicated by the *Bayes factor,*

$$B_{ij} \equiv p(\mathbf{X},\mathbf{y}|M_i)/p(\mathbf{X}, \mathbf{y}|M_j), \qquad [9]$$

the ratio of the marginal likelihoods for model M_i and model M_j. Since, by Equation [7],

$$\pi_j \propto p(\mathbf{X}, \mathbf{y}|M_j)\, \pi_j^0 \qquad [10]$$

for model M_j and

$$\pi_i \propto p(\mathbf{X}, \mathbf{y}|M_i)\, \pi_i^0, \qquad [11]$$

[5]Notice that if the regression parameter estimate $\mathbf{b}_j$ was identical for each model, the weighted sum of squared deviations from the weighted average $\bar{\mathbf{b}}$ in the second term on the right-side of Equation [6] would be exactly zero, leaving the unconditional posterior variance equal to the weighted average of the conditional posterior variances in the first term.

for model M_i, the Bayes factor transforms the prior model odds ratio π_i^0/π_j^0 into a posterior model odds ratio π_i/π_j:

$$\pi_i/\pi_j = B_{ij}\, \pi_i^0/\pi_j^0. \qquad [12]$$

When B_{ij} exceeds one, the data favor model M_i over model M_j and the posterior odds ratio exceeds the prior odds ratio; when B_{ij} is less than one the reverse is true.

In general, calculating a Bayes factor requires calculating the marginal likelihood for each model, which in turn requires specifying a prior distribution $p(\boldsymbol{\theta}_j|M_j)$ for the parameters in each model M_j, evaluating the conditional likelihood $p(\mathbf{X}, \mathbf{y}|\boldsymbol{\theta}_j, M_j)$ for each $\boldsymbol{\theta}_j$, and integrating the product of the conditional likelihood and the prior. Unfortunately, the practical difficulties involved in constructing an appropriate model-specific prior distribution for the parameters in each model M_j and analytically or numerically integrating the resulting multidimensional marginal likelihood are daunting (though recent advances in computational statistics offer some promise of progress on the latter front).

An additional difficulty arises in considering Bayes factors for nested models, since the conditional likelihood of a more general model in which a given parameter is unconstrained must always exceed that of an otherwise similar but simpler model in which the parameter is constrained.[6] Thus, if we want to assign positive probability to relatively simple models, we must adjust the marginal likelihoods in some way that rewards parsimony, in effect specifying a proper prior distribution $p(\boldsymbol{\theta}_j|M_j)$ in Equation [8] that penalizes the estimation of additional parameters.

Raftery (1995) and Kass and Wasserman (1995) have proposed a simple method for approximating Bayes factors that minimizes computational difficulties while allowing for direct comparisons of nested models. They show that, with a proper prior over the model parameters equivalent to a single typical observation, the log of the marginal likelihood is approximately equal to the log of the maximized likelihood minus a penalty proportional to the number of parameters being estimated.[7]

[6]Leaving the additional parameter unconstrained will always increase the likelihood of the data at least slightly, in much the same way that adding an additional variable to a regression equation will always increase the value of the (unadjusted) R^2 statistic at least slightly.

[7]More precisely, the required prior distribution for $\boldsymbol{\theta}_j$ is multivariate normal with mean vector $\hat{\boldsymbol{\theta}}_j$ and covariance matrix $\mathbf{i}_j^{-1}$, where $\hat{\boldsymbol{\theta}}_j$ is the vector of maximum likelihood (here, ordinary least squares) estimates of $\boldsymbol{\theta}_j$ and $\mathbf{i}_j$ is the expected Fisher information matrix for one observation ($\mathbf{i}_j \approx \mathbf{A}_j/\mathrm{N}$, where $-\mathbf{A}_j$ is the Hessian matrix of second partial derivatives of the conditional log-likelihood evaluated at $\hat{\boldsymbol{\theta}}_j$). Then the log of the marginal likelihood is equal to the log of the maximized likelihood minus $\ln(\mathrm{N})\ (\mathrm{K}_j/2)$ plus an approximation

The penalty for additional parameters implied by Raftery (1995) and Kass and Wasserman's (1995) approximation to the marginal likelihood turns out to be precisely the same penalty implied by the *Bayesian Information Criterion* (BIC) proposed by Schwarz (1978) as a model selection criterion. In particular (Raftery 1995), the BIC for model M_j is an approximate function of B_{j0}, the Bayes factor for a comparison of model M_j and a baseline model M_0 including only an intercept:

$$\mathrm{BIC}(M_j) \approx -2\ \ln(B_{j0}). \tag{13}$$

In the case of a linear regression model with normal errors, the BIC is

$$\mathrm{BIC}(M_j) = \mathrm{N}\ \ln(1 - \mathrm{R}_j^2) + \ln(\mathrm{N})\ (\mathrm{K}_j - 1), \tag{14}$$

where R_j^2 is the regression R^2 statistic for model M_j, K_j is the number of parameters estimated in model M_j, including the intercept, and smaller (more negative) values of BIC indicate larger Bayes factors and, hence, higher posterior model probabilities (Kass and Wasserman 1995).[8] Substituting expression [14] into expression [13] and rearranging, the Bayes factor for model M_j relative to the baseline model M_0 is approximately

$$B_{j0} \approx \exp\{(-\mathrm{N}/2)\ \ln(1 - \mathrm{R}_j^2) - \ln(\mathrm{N})\ ((\mathrm{K}_j - 1)/2)\}, \tag{15}$$

which can be easily computed from standard regression output.

Once we have calculated (or approximated) the Bayes factor B_{j0} for each of several alternative models, we can solve for the model posterior probability for any particular model M_i as a function of the complete set of model prior probabilities and Bayes factors by repeated application of Equation [12]:

$$\pi_i = B_{i0}\ \pi_i^0 \Big/ \sum_j B_{j0}\ \pi_j^0.^9 \tag{16}$$

error of order $1/\sqrt{\mathrm{N}}$, where K_j is the number of parameters estimated in model M_j and N is the sample size.

[8]The Bayesian Information Criterion defined in Equation [14] is essentially the Schwarz criterion rescaled so that BIC (M_0) equals zero; thus, for comparisons of alternative models with the same dependent variable and sample size, as here, BIC and the Schwarz criterion are equivalent model selection criteria. The BIC tends to prefer somewhat more parsimonious models than do most of the more familiar model selection criteria, such as the adjusted R^2 statistic or the Akaike Information Criterion (Judge et al. 1985, 862–75; Beck and Katz 1992).

[9]In the special case of uniform model priors ($\pi_1^0 = \ldots = \pi_j^0 = \ldots = \pi_J^0 = 1/J$), the posterior probability for each model is simply proportional to the corresponding Bayes factor.

The model posterior probabilities can be used in turn to weight the regression parameter estimates and variances from the various alternative models in order to calculate the model mixture posterior means and variances using Equations [5] and [6]. An example of the required calculations is presented in note 13 below.

The approach described here, in which the relevant Bayes factors are approximated on the basis of the corresponding BICs, is only one of a variety of approaches to calculating the required model posterior probabilities. A more complicated but putatively more accurate approximation applicable to generalized linear models is described by Raftery (1994), Kass and Raftery (1995) and Raftery (1995) provide useful overviews.[10]

Examples

In this section I apply the model-averaging procedure outlined in the previous section to two prominent political science data analyses. My aims are to illustrate the calculations involved in the model-averaging procedure, to introduce some alternative reference priors that might usefully represent different assumptions about the *a priori* plausibility of alternative models, and to provide some raw material for an evaluation of the potential fruitfulness of model averaging as a technique for statistical inference under model uncertainty.

Politics and Economic Growth in OECD

My first example is based on a controversy between Lange and Garrett (1985, 1987) and Jackman (1987) regarding the impact of the organizational and political strength of labor on relative economic growth rates in advanced industrial democracies in the late 1970s. Lange and Garrett hypothesized that both corporatist systems (with well-organized labor movements and predominantly leftist governments) and free market systems (with poorly-organized labor movements and predominantly rightist gov-

---

More generally, the model posterior probability is proportional to the product of the corresponding Bayes factor and the model prior.

[10]"Putatively" in that the practical limitations of the various approximations are not yet clear. Kass and Raftery (1995, 778) suggested on the basis of some simulation evidence that sample sizes larger than 20 times the number of parameters to be estimated are "large enough for the method to work well," while sample sizes smaller than five times the number of parameters to be estimated are "worrisome." I have replicated Western's (1996) analysis of welfare state decommodification, which has six parameters to be estimated from 18 observations, using the simple BIC approximation described here, and obtained results identical within rounding error to those produced by Western using Raftery's (1994) more sophisticated approximation.

Table 1. Regression Analysis of Economic Growth in OECD (N = 15)

	Model 1	Model 2	Model 3	Model 4	Model 5
Intercept	.685	.568	.650	.521	.558
	(.255)	(.099)	(.096)	(.074)	(.092)
Labor Organization Index	−.214	−.070	−.038	−.021	−.107
	(.199)	(.064)	(.058)	(.049)	(.050)
Left Party Vote	−.671				
	(.696)				
Labor × Left Vote	.631				
	(.458)				
Left Party Cabinet Portfolios		−.741	−.653	−.370	−.397
		(.309)	(.275)	(.252)	(.261)
Labor × Left Portfolios		.366	.296	.147	.156
		(.125)	(.115)	(.113)	(.117)
Dependence on Imported Oil			−.245		−.088
			(.119)		(.124)
Norway Dummy				.475	.401
				(.145)	(.182)
Standard error of regression	.172	.139	.122	.101	.104
R^2	.235	.501	.649	.759	.772
Adjusted R^2	.027	.365	.509	.663	.645
BIC	4.10	−2.31	−4.88	−10.52	−8.62

ernments) would be capable of weathering the economic shocks of the period, but that systems with mismatches between labor organization and political power—either well-organized labor movements with predominantly rightist governments or poorly-organized labor movements with predominantly leftist governments—would be less capable of sustaining economic growth. They tested their hypothesis using data on relative growth rates in 15 OECD countries, and interpreted their empirical results as "strongly confirm[ing]" their hypothesis (Lange and Garrett 1985, 821).

The first three regression models reported in Table 1 are identical to three of the 10 models presented by Lange and Garrett (1985, Table 3).[11] (I omit Lange and Garrett's other seven models because they do not include the interaction term of central theoretical interest, and seem to be included

[11]The parameter estimates in Table 2 differ from those reported by Lange and Garrett (1985, Table 3) and Jackman (1987, Table 1) because I have not standardized the data for Left Party Vote and Left Party Cabinet Portfolios as Lange and Garrett did (Jackman 1987, note 3). They also seem to have standardized the data for Dependence on Imported Oil, which I have reproduced from their cited source.

simply to demonstrate the inadequacy of additive models.[12]) Model 1 allows for an interaction between the organizational strength of labor and the average vote share of left parties; in Model 2 the share of cabinet portfolios held by left parties is substituted for the average vote share of left parties as a measure of the political strength of labor. Model 3 adds to the specification in Model 2 (which fits the data much better than the specification in Model 1) a measure of dependence on imported oil, presumably in order to capture the effect of the most significant exogenous economic shock in the period under study. Model 3 in turn fits the data significantly better than Model 2 does, and provides the basis for Lange and Garrett's subsequent interpretation and discussion.

Jackman (1987) criticized Lange and Garrett's analysis on several grounds, arguing most notably that their results were dominated by a single influential observation, Norway, which had a very favorable corporatist structure from Lange and Garrett's theoretical perspective, a great deal of oil, and a very high relative growth rate. His reanalysis adding a dummy variable for Norway to Model 2 (Jackman 1987, Table 1) is reproduced as Model 4 in Table 1. Model 4 fits the data significantly better than any of Lange and Garrett's models do, and provides the basis for Jackman's subsequent interpretation and discussion.

For completeness, Model 5 in Table 1 combines the features of Lange and Garrett's preferred model (including dependence on imported oil) and Jackman's preferred model (including the dummy variable for Norway). Model 5 fits the data slightly less well than Model 4 does, with the Norway dummy variable but not the oil variable producing a large and "statistically significant" parameter estimate.

It seems clear that any assessment of the evidence in Table 1 will be quite sensitive to the relative weight we attach to the various alternative models. At one extreme, Lange and Garrett's model, reproduced in the first column of Table 2, seems to provide strong support for the hypothesized interaction between labor organization and political strength, with sizable coefficients and hefty t-statistics (2.6 and 2.4, respectively) on the interaction term and the baseline effect of Left Party Cabinet Portfolios. At the other extreme, Jackman's model, reproduced in the second column of Table 2, seems to provide much less support for the hypothesized interaction, with the relevant coefficients only about half as large and t-statistics of 1.3 and 1.5, respectively.

The third column of Table 2 presents Bayesian mixture posterior coefficients based on uniform model priors, which assign equal prior probability

[12]Including these additive models would leave the results of the overall analysis virtually unchanged, since none of them fits the data well.

Table 2. Alternative Bayesian Mixture Posteriors for Economic Growth in OECD

	Lange and Garrett (Model 3)	Jackman (Model 4)	Uniform Model Priors	Dummy-Resistant Model Priors
Intercept	.650	.521	.536	.577
	(.096)	(.074)	(.086)	(.108)
Labor Organization	−.038	−.021	−.021	−.032
Index	(.058)	(.049)	(.051)	(.059)
Left Party Vote			−.000	−.003
			(.021)	(.061)
Labor × Left Vote			.000	.002
			(.017)	(.049)
Left Party Cabinet	−.653	−.370	−.393	−.508
Portfolios	(.275)	(.252)	(.265)	(.307)
Labor × Left Port-	.296	.147	.158	.222
folios	(.115)	(.113)	(.120)	(.142)
Dependence on Im-	−.245		−.033	−.100
ported Oil	(.119)		(0.90)	(.140)
Norway Dummy		.475	.431	.248
		(.145)	(.185)	(.255)
Posterior probabilities				
Model 1	.0000	.0000	.0005	.0040
Model 2	.0000	.0000	.0113	.0974
Model 3	1.0000	.0000	.0407	.3521
Model 4	.0000	1.0000	.6829	.3939
Model 5	.0000	.0000	.2647	.1527

to each of the five models presented in Table 1. The corresponding posterior model probabilities, reported at the bottom of the column, are markedly nonuniform, reflecting the relative goodness of fit of the models including the Norway dummy variable. The model estimated by Jackman (Model 4) gets almost 70% of the posterior probability, and the model adding the Norway dummy variable to Lange and Garrett's preferred specification (Model 5) gets about 25%, while all of Lange and Garrett's models together get only a little more than 5% of the posterior weight. As a result, the Bayesian mixture coefficients are generally similar to the parameter estimates from Model 4. However, by comparison with the standard errors from Model 4, the posterior mixture standard errors are from 4% larger (in the case of labor organization) to 28% larger (in the case of the Norway dummy variable), reflecting the extent to which uncertainty about the correct model specification should detract from the apparent precision of Jackman's results.[13]

While the posterior mixture coefficients and standard errors in the third column of Table 2 seem to provide a more reasonable summary of the data than either Lange and Garrett's or Jackman's preferred results taken alone, they remain problematic in at least two respects. First, they do almost nothing to reflect Lange and Garrett's original uncertainty about whether vote shares or cabinet portfolios are the more appropriate measure of the political strength of labor, because the specification using vote shares (Model 1) is swamped by the subsequent specifications using cabinet portfolios. While the cabinet portfolio measure seems clearly preferable on theoretical as well as empirical grounds, it would seem appropriate to attach some additional uncertainty to the results of the analysis to reflect the fact that Lange and Garrett's original hypothesis about the interactive effect of labor organization and political strength had two separate chances to fit the data, only one of which actually panned out. I return to this issue in the Discussion section below.

Second, and more problematically, the Bayesian mixture posterior with uniform model priors is dominated by the results from Model 4 and Model 5, both of which include Jackman's dummy variable for Norway. But since the inclusion of the Norway dummy variable was prompted by preliminary data analysis rather than by *a priori* theoretical considerations, it would seem appropriate to discount the results from the models including the Norway dummy variable, rather than pretending that those models were as plausible *a priori* as the models without dummy variables. One reasonable way to do that is to specify ''dummy-resistant model priors'' that attach relatively less prior probability to models including a country-specific dummy variable.

[13]The approximate Bayes factor for each regression model in Table 1 can be calculated (within rounding error) from the reported R^2 statistic (and the number of estimated parameters shown in the table) using expression [15], or from the reported BIC statistic using expression [13]. The approximate Bayes factors for Models 1 through 5, respectively, are .129, 3.171, 11.461, 192.327, and 74.543, the sum of the Bayes factors for all five models is 281.631, and the model posterior probabilities in the third column of Table 2 (derived from uniform model priors) are the corresponding Bayes factors divided by the sum of the Bayes factors as in Equation [16]. The mean and standard deviation of the mixture posterior for each coefficient in the third column of Table 2 can be calculated from the regression parameter estimates and standard errors in Table 1 and the corresponding model posterior probabilities using Equations [5] and [6]. For example, the mean of the mixture posterior for the Labor $\times$ Left Portfolios interaction term is $.0005 \times 0$ (from Model 1) + $.0113 \times .366$ (from Model 2) + $.0407 \times .296$ (from Model 3) + $.6829 \times .147$ (from Model 4) + $.2647 \times .156$ (from Model 5) = .158.

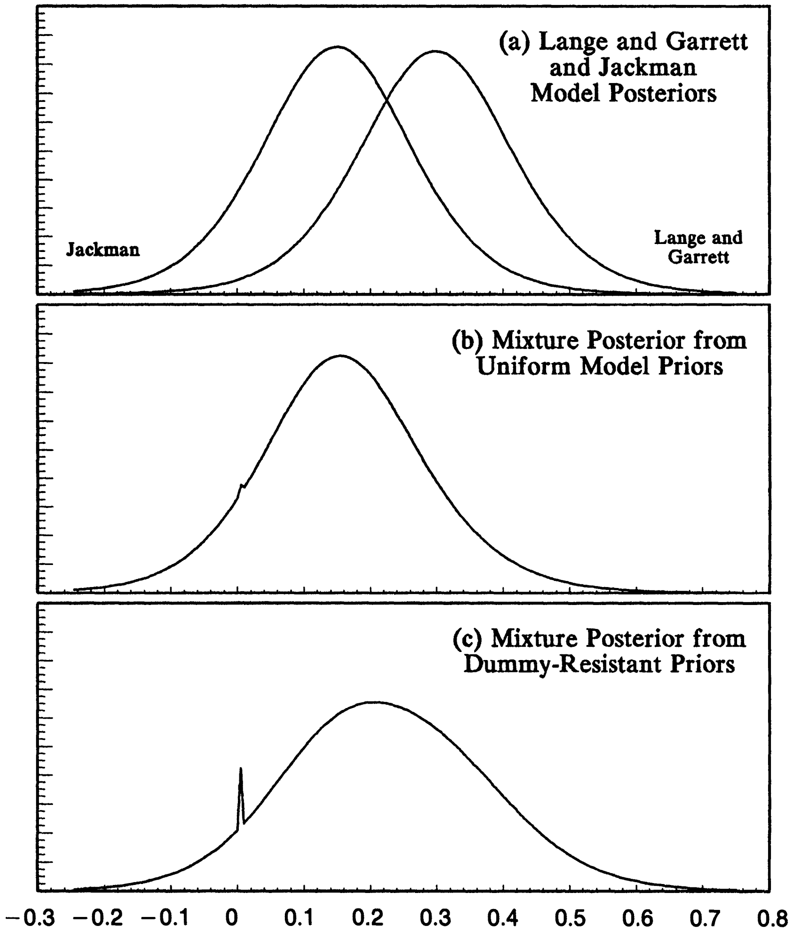
Since in the present case there are 15 countries in the data set, each equally likely *a priori* to be an outlier, it seems reasonable to attach one-fifteenth as much prior probability to the models including the Norway dummy variable as to the models including only combinations of Lange and Garrett's original variables. The resulting "dummy-resistant" model priors of .3191 for Models 1, 2, and 3 and .0213 for Models 4 and 5 reflect considerable skepticism about the plausibility of models with *ad hoc* variables added to account for single outliers, while nevertheless allowing some scope for *a priori* skepticism to be overcome by clear signals from the data.[14]

The Bayesian mixture posteriors produced by averaging the results from the five models in Table 1 using these dummy-resistant model priors instead of uniform model priors are presented in the last column of Table 2. Here, Lange and Garrett's preferred model (Model 3) and Jackman's preferred model (Model 4) have approximately equal posterior probabilities, and the Bayesian mixture posterior coefficients fall roughly halfway between these two extremes. Neither Lange and Garrett's oil variable nor Jackman's Norway dummy variable approach "statistical significance" (with t-statistics of 0.7 and 1.0, respectively), while the coefficients for Lange and Garrett's Left Cabinet Portfolios and Labor × Left Portfolios interaction variables are 30 to 40% larger than with uniform model priors (though still 20 to 25% smaller than in Lange and Garrett's preferred model), with t-statistics of about 1.6.[15]

[14]Obviously, there is no reason why the prior weights for models with case-specific dummy variables must be proportional to 1/N. If Norway was known in advance to be a notorious outlier in regressions of this sort, even after controlling for the effects of North Sea oil, we might well attach greater prior plausibility to models with a Norway dummy variable. The important point is that we should do so consciously before analyzing the data, and should recognize the implications of that assessment for our subsequent inferences *even if the Norway dummy variable turns out to be unnecessary.* It is simply too easy to justify *ad hoc* model specifications after the fact, forgetting in the process the various alternative specifications that might also have been justified had they turned out to work well by one criterion or another. Choosing a prior weight proportional to 1/N is a crude but potentially effective way to impose some discipline upon such *post hoc* rationalizations.

[15]I report t-ratios for the model mixture posterior coefficients for descriptive purposes only. The t-distribution will not be a good approximation for the actual posterior distribution of any coefficient set to zero with certainty in models which get significant posterior weight, since the posterior will be a mixture of a t-distribution and a spike at zero; two mild examples are shown in Figure 1. An analyst intent upon calculating the posterior probability that a parameter is greater than (less than) zero could do so by using a t-ratio based on the mean and standard deviation of the conditional posterior mixture distribution computed from all models in which the relevant parameter was *not* set to zero with certainty, and multiplying the resulting conditional posterior probability by the total posterior probability of all such models.

Figure 1. Alternative Posteriors for Labor × Left Portfolios Interaction.



The three panels of Figure 1 provide a graphical comparison of the alternative posterior distributions summarized in Table 2 for Lange and Garrett's Labor × Left Portfolios interaction variable. Panel (a) shows the posterior distributions for Model 3 (Lange and Garrett's preferred model, on the right) and Model 4 (Jackman's preferred model, on the left) from Table 1. Panel (b) shows the mixture posterior corresponding to uniform priors over the five models in Table 1. Panel (c) shows the mixture posterior corresponding to the dummy-resistant model priors in which each of the models including the Norway dummy variable gets only one-fifteenth as much prior weight as each of the other models.

There is a very small spike at zero in the posterior distribution from the uniform model priors and a more perceptible spike at zero in the posterior distribution from the dummy-resistant priors, in each case reflecting the posterior weight attached to Model 1 from Table 1, where the relevant variable was omitted (i.e., set to zero with certainty) in favor of the Labor × Left Vote variable.[16] In other respects, the posterior distribution from the uniform model priors resembles the posterior distribution for the Jackman model, whereas the posterior distribution for the dummy-resistant model priors is a compromise between the posterior distributions for the Lange and Garrett and Jackman models, with a standard deviation about 25% larger than either due to the posterior specification uncertainty.[17]

Generally speaking, the dummy-resistant posterior coefficients reported in the fourth column of Table 2 and illustrated in panel (c) of Figure 1 seem to me to provide the best summary of the import of Lange and Garrett's data. They provide some real support for the hypothesized interaction between labor organization and political power, but that support is significantly tempered by specification uncertainty, and especially by the sensitivity of the results from this small data set to the influence of a single problematic observation. As Lange and Garrett (1987, 268, 272) themselves put it, with appropriate caution, "Jackman's rejection of our hypotheses on the basis of the exclusion of Norway is not conclusively supported," and some "evidence remains that domestic political structures had an impact on the relative decline of growth performance."[18]

[16]Obviously, the posterior mixture distributions for the Norway and oil dependence variables—which are omitted from models whose posterior model weights greatly exceed that of Model 1—have much more pronounced spikes at the origin than appear in Figure 1.

[17]By comparison, the standard deviation of the posterior distribution for the main effect of Left Party Cabinet Portfolios with dummy-resistant model priors is 12% larger than in Lange and Garrett's model and 22% larger than in Jackman's model. The standard deviation of the posterior distribution for the Oil Dependence coefficient is 18% larger than in Lange and Garrett's model, and the standard deviation of the posterior distribution for the Norway coefficient is 76% larger than in Jackman's model (and 38% larger than with uniform model priors). In each case, the increased standard error of the parameter estimate reflects the impact of specification uncertainty unacknowledged in each of the regressions considered separately.

[18]Lange and Garrett's data are presented in Table A1 in the Appendix. Beck and Katz (1992) used the same data, but without the oil dependence and Norway variables, to illustrate model selection by cross-validation. Unlike the Bayesian Information Criterion applied here, the cross-validation criterion prefers the baseline model M_0 containing only an intercept to Model 2 in Table 1. It is hard for me to see why a social scientist—as distinct from a goodness-of-fit statistic—would prefer a null model to one with real theoretical content. But

Public Opinion and Policy in the American States

My second example is based on Erikson, Wright, and McIver's (1993) analysis of the impact of public opinion on policy outcomes in the American states. Erikson, Wright, and McIver derived an index of state policy liberalism from detailed data on state policies in eight issue areas, and an index of state public opinion from an accumulation of CBS News/*New York Times* polls. They reported results from a variety of models in which policy liberalism was regressed on public opinion and other explanatory variables, including the relative liberalism of each state's political elites and the proportion of Democrats in each state's legislature. Several of these models included explanatory variables derived from Elazar's (1972) classification of state political cultures as "individualistic," "moralistic," or "traditionalistic," either alone or in interaction with state opinion or elite liberalism.[19]

The six models reported in Table 3 are identical to six of the eight models of state policy liberalism reported by Erikson, Wright, and McIver (1993, 159, 170).[20] They range from a very simple specification with State

---

for present purposes it is more important to note that the model-averaging approach described here dispenses with the necessity of settling on *any* single "best" model when there are a variety of *a priori* plausible alternatives. Further analyses using similar models with a pooled time-series cross-section design, albeit for roughly the same countries and time period, were reported by Alvarez, Garrett, and Lange (1991) and Beck et al. (1993).

[19]The relevant data are presented in Table A2 in the Appendix. Erikson, Wright, and McIver (1993) presented detailed information regarding conceptualization and measurement of the various variables (especially in Chapters 2, 5, and 7) and auxiliary analyses documenting the additional impact of State Opinion on Legislative Liberalism and Democratic Legislative Strength (Tables 7.4 and 7.5, respectively).

[20]Models 1 through 6 in my Table 3 correspond to Models 2, 3, and 4 in Erikson, Wright, and McIver's Table 7.1 (1993, 159) and Models 1, 2, and 3 in their Table 7.6 (1993, 170), respectively. I omit the one model in Erikson, Wright, and McIver's analysis that does not include state opinion as an explanatory variable. I also omit one more complicated model with additional interaction terms; under any of the model priors presented in Table 4, that model would have a posterior probability of .0005 or less, leaving the overall results virtually unchanged. The results presented in Table 3 differ from Erikson, Wright, and McIver's for three reasons. First, I use the raw values of the explanatory variables presented in their tables rather than standardizing all the continuous variables as they did in their regression analyses. Second, I omit Nebraska from the analyses for Models 1, 2, and 3 to preserve comparability in the set of observations across models. (Erikson, Wright, and McIver omitted Nebraska from their regressions including the Democratic Legislative Strength variable because the state has a nonpartisan legislature, but included Nebraska in their earlier regressions.) Third, there may be minor discrepancies due to my use of Erikson, Wright, and McIver's published data with whatever rounding errors they produce (and my independent reconstruction of the Democratic Legislative Strength measure, for which Erikson, Wright, and McIver did not publish their data). None of these differences turns out to be consequential, since the results presented here essentially match Erikson, Wright, and McIver's, allowing for the differences in scale produced by my use of unstandardized data.

Table 3. Regression Analysis of State Policy Liberalism (N = 46)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	1.640	1.483	1.854	1.741	1.216	3.295
	(.185)	(.160)	(.224)	(.310)	(.372)	(.891)
State Opinion	.1137	.0955	.1383	.0599	.0570	.1547
	(.0114)	(.0113)	(.0217)	(.0166)	(.0183)	(.0365)
Legislative Liberalism				.2834	.2112	.0777
				(.0739)	(.0826)	(.0979)
Democratic Legislative Strength				−1.708	−.591	−2.368
				(.376)	(.608)	(1.146)
Elazar Traditionalism		−.581	−1.507		−.545	−2.815
		(.208)	(.533)		(.248)	(1.063)
Traditionalism × State Opinion			−.0717			−.0988
			(.0327)			(.0394)
Traditionalism × Dem Leg Strength						1.960
						(1.330)
Elazar Moralism		.258	−.244		.069	−3.547
		(.181)	(.326)		(.192)	(1.423)
Moralism × State Opinion			−.0519			−.1138
			(.0262)			(.0387)
Moralism × Dem Leg Strength						4.732
						(2.083)
Standard error of regression	.563	.456	.437	.451	.432	.406
R^2	.693	.807	.831	.811	.835	.869
Adjusted R^2	.686	.794	.810	.798	.815	.837
BIC	−50.48	−64.30	−62.74	−65.26	−63.86	−59.15

Opinion as the only explanatory variable to much more complex specifications, including one with distinct culture-specific effects for State Opinion and Democratic Legislative Strength and a total of 10 parameters to be estimated from the 46 observations.[21] All of these models fit the data very well, with the adjusted R^2 statistics for all but the first ranging from .79 to .84.

Erikson, Wright, and McIver (1993, 169) ''focus mainly on the 'best' equation'' corresponding to Model 6 in Table 3, which is the most complex of the alternative specifications. Eight of the 10 coefficients estimated in that model have *t*-statistics of 2.0 or greater, and the average *t*-statistic for the six coefficients capturing the distinctive effects of state political cultures as classified by Elazar is 2.4. The results suggest that individualistic states have much more liberal baseline policies than traditionalistic or moralistic states, and are also much more sensitive to variations in state opinion. On the other hand, having more Democrats in the state legislature actually seems to produce more conservative policies in individualistic states, but more liberal policies in moralistic states (with no effect in traditionalistic states). Erikson, Wright, and McIver (1993, 173) interpreted these results as reflecting a systematic difference between ''pragmatic'' and ''ideological'' political systems:

> The startlingly large differences in the coefficients for moralistic and individualistic states can be attributed to the difference between pragmatic politicians in individualistic states and the programmatic politicians in moralistic states. In moralistic states, the relative partisan division of legislatures has important policy consequences, while legislatures are less responsive to the direct input of public opinion. In individualistic states, the party division has less easily discernible policy consequences, but the pragmatic legislatures respond strongly to state opinion.

Most of the culture-specific variables also have ''statistically significant'' coefficients in the other models where they appear in Table 3. But the magnitudes of these coefficients vary considerably among the various specifications, as do the estimates of the baseline effect of state opinion. Thus, it is by no means obvious how well Erikson, Wright, and McIver's interpretation of the results from their ''best'' model would hold up in an analysis which averaged the results produced by the various alternative models.

[21] I follow Erikson, Wright, and McIver in omitting Alaska, Hawaii, and Nebraska from the analysis due to missing data, and Nevada due to an implausibly liberal State Opinion score.

Table 4 presents the means and standard deviations of the Bayesian mixture posteriors for each coefficient, as well as the model posterior probabilities, for a variety of alternative model priors. By contrast with the example presented in Table 1, none of the alternative models presented in Table 3 clearly dominates the others by the Bayesian Information Criterion. This fact is reflected in the model posterior probabilities reported at the bottom of Table 4. With uniform model priors (in the first column of Table 4), Model 4 gets 41% of the posterior weight, while Model 2, Model 3, and Model 5 each get between 12 and 25%; Erikson, Wright, and McIver's "best" model, Model 6, gets only 2% of the posterior weight—an indication of the extent to which the Bayesian Information Criterion discounts good fits achieved at the cost of estimating additional parameters.[22]

Still focusing on the case of uniform model priors—the first column of Table 4—the posterior mixture coefficient for State Opinion produced by averaging the six distinct parameter estimates in Table 3 has a mean of .0792, suggesting that the difference in public opinion between the five most liberal states (Massachusetts, Rhode Island, New York, New Jersey, and Connecticut) and the five most conservative states (Utah, Idaho, Oklahoma, North Dakota, and Mississippi) probably produced public policies that differed by almost two standard deviations on the state policy liberalism scale. This result clearly confirms Erikson, Wright, and McIver's central hypothesis concerning the impact of public opinion on state policy outcomes. Even here, though, the impact of specification uncertainty is substantial: the magnitude of the estimated effect is only half as large as in Erikson, Wright, and McIver's single "best" specification, Model 6, and the corresponding t-statistic of 2.4 is a good deal less impressive than the t-statistic of 4.2 for the same variable in Model 6, much less the t-statistics of 6, 8, and 10 in some of the other models in Table 3.

Moreover, Erikson, Wright, and McIver's other explanatory variables fare much worse in the model averaging process. The magnitudes of the mixture coefficients for all but one—Legislative Liberalism—are much reduced (by an average of 90%) relative to the estimates in Erikson, Wright, and McIver's "best" specification. And none even approaches conventional significance levels (the largest t-statistic is 1.1, and the average is

[22]More generally, the extent to which the Bayesian Information Criterion (and hence the model posterior probabilities) reflect a preference for parsimony is evident from a comparison of the regression results in Table 3 and the model posterior probabilities in Table 4. With uniform model priors, Model 4 gets twice as much posterior weight as Model 5, 3.5 times as much as Model 3, and more than 20 times as much as Model 5—despite having a larger regression standard error and a smaller adjusted R^2 statistic than any of these alternative models—because it has from two to six fewer parameters to be estimated.

Table 4. Alternative Bayesian Mixture Posteriors for State Policy Liberalism

	Uniform Model Priors	Slightly Search-Resistant Model Priors	Moderately Search-Resistant Model Priors	Strongly Search-Resistant Model Priors
Intercept	1.612	1.610	1.606	1.601
	(.446)	(.418)	(.390)	(.361)
State Opinion	.0792	.0810	.0831	.0855
	(.0336)	(.0332)	(.0327)	(.0319)
Legislative Liberalism	.1603	.1486	.1345	.1180
	(.1419)	(.1435)	(.1440)	(.1425)
Democratic Legislative Strength	−.864	−.804	−.733	−.648
	(.885)	(.880)	(.869)	(.850)
Elazar Traditionalism	−.486	−.494	−.508	−.525
	(.635)	(.616)	(.597)	(.576)
Traditionalism × State Opinion	−.0102	−.0103	−.0103	−.0103
	(.0288)	(.0287)	(.0286)	(.0284)
Traditionalism × Dem Leg Strength	.038	.029	.021	.015
	(.326)	(.287)	(.247)	(.206)
Elazar Moralism	−.017	.006	.030	.056
	(.577)	(.523)	(.470)	(.419)
Moralism × State Opinion	−.0082	−.0081	−.0079	−.0078
	(.0245)	(.0237)	(.0230)	(.0223)
Moralism × Dem Leg Strength	.091	.070	.052	.036
	(.710)	(.626)	(.538)	(.448)
Posterior probabilities				
Model 1	.0003	.0003	.0004	.0006
Model 2	.2527	.2976	.3513	.4151
Model 3	.1158	.1228	.1289	.1332
Model 4	.4091	.3902	.3640	.3292
Model 5	.2029	.1742	.1445	.1143
Model 6	.0192	.0148	.0109	.0076

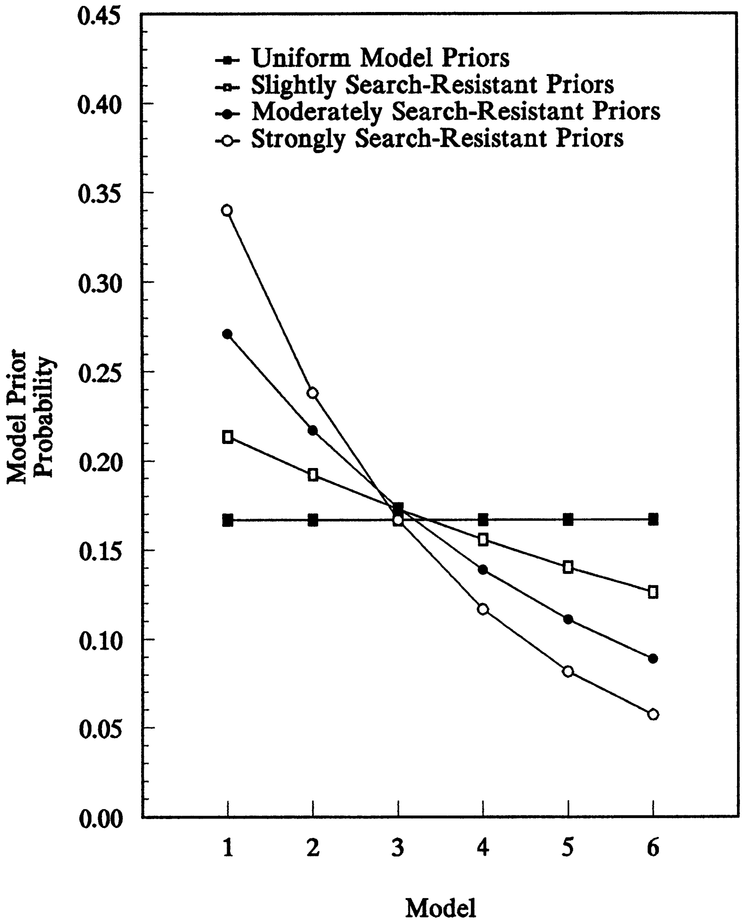
0.5), despite the fact that their t-statistics in the separate models where they appear in Table 3 average almost 2.0.

The contrast between these mixture posteriors and the original regression results in Table 3 is striking. Might the discrepancies be attributable to some peculiarity of the uniform model priors from which these mixture posteriors are derived? In order to test the sensitivity of the mixture posteriors to the assumptions embodied in the model priors, it may be useful to investigate the implications of some alternative sets of model priors. For example, the uniform model priors attach no significance to the order in which the various regression models were estimated. While there is no logical connection between the relative prior plausibility of alternative models and the order in which they were actually estimated, it seems reasonable to suppose that data analysts typically begin by estimating the model they consider most plausible *a priori,* proceeding to relatively less plausible models until they reach some acceptable stopping point.[23] If that is the case, then the prior probabilities attached to the various alternative models should reflect their place in the sequence, with later specifications receiving less prior weight than earlier specifications.

One rough way to capture the potential inferential significance of sequential search strategies is to adopt a "search-resistant prior" that attaches smoothly declining prior weights to each model in a sequence of alternative possibilities. Rather than simply penalizing complex models for their relative lack of parsimony, as the BIC and most other model selection criteria already do, a search-resistant prior penalizes the *process* of sequential model estimation itself, whether it results in more or less parsimonious models. The penalty reflects the fact that latter models in the sequence are more likely to capitalize on chance, since they are typically formulated, in part, on the basis of the results produced by earlier models in the sequence. It is important to note, however, that a search-resistant prior only "penalizes" later models relative to earlier ones; it does not alter the basic logic of model averaging, or make the posterior results as a whole any more or less uncertain. The contribution of specification uncertainty to overall uncertainty will still depend upon how much the results vary across plausible specifications—especially plausible specifications that fit the data about equally well.

[23]The psychology of stopping rules is complicated in its own right. Data analysts seem inclined to end their specification searches when they get "pleasing" regression results (big t-statistics, "correct" signs, or whatever), but their standards for what is "pleasing" (or at least "acceptable") may vary as they approach data exhaustion, physical exhaustion, or exhaustion of imagination. The fact remains that the prospect of stopping sooner rather than later makes it rational to investigate more plausible specifications (that is, those more likely to produce "pleasing" results) before less plausible specifications.

Figure 2. Alternative Model Priors for Specification Searches.



One possible family of search-resistant model priors is shown in Figure 2. The "slightly search-resistant priors" shown in Figure 2 attach a prior probability proportional to $.9^j$ to model M_j, so that each model in the sequence gets 10% less prior weight than the one before. The "moderately search-resistant priors" shown in Figure 2 attach a prior probability proportional to $.8^j$ to model M_j, discounting each subsequent model by 20% relative to the one before, while the "strongly search-resistant priors" attach a prior probability proportional to $.7^j$ to model M_j, discounting each subsequent model by 30% relative to the one before.

Simple reference priors of this sort can never be expected to capture all the nuances of a data analyst's real prior beliefs. Nor, at the opposite extreme, are they likely to make sense of a sequence of models generated by a largely or wholly mechanical model selection procedure such as forward or backward stepwise regression, where the selection and order of estimation of alternative models is simply a complicated function of their goodness of fit and an arbitrary starting point, rather than a reflection of considered judgment by the data analyst. However, in many cases where a sequence of models does reflect at least the rough contours of a sophisticated data analyst's substantive judgment, a simple reference prior of the sort proposed here may be quite useful for exploring the sensitivity of statistical inferences to general features of the specification search strategy that generated a given set of statistical results.

In Erikson, Wright, and McIver's case, uniform model priors attach a prior probability of .1667 to each of the six models presented in Table 3. By contrast, the prior probabilities with slightly search-resistant priors range from .2134 for Model 1 down to .1260 for Model 6, making Model 1 about 70% more probable than Model 6 *a priori.* The prior probabilities with strongly search-resistant priors range from .3400 for Model 1 down to .0571 for Model 6, making Model 1 about six times as probable as Model 6 *a priori.* Thus, these alternative reference priors allow us to explore a considerable range of assumptions regarding the extent to which later models should be discounted as less plausible *a priori* than earlier models in the sequence.

Of course, none of these alternative model priors will perfectly reflect Erikson, Wright, and McIver's own prior beliefs about the plausibility of their various models, or anyone else's. Nevertheless, if all of these various priors produce mixture posteriors similar to those produced by the uniform model priors, it will be hard to resist the conclusion that Erikson, Wright, and McIver's informal synthesis of their various results overstated the weight of their evidence in support of Elazar's classification of state political cultures, given the specification uncertainty inherent in their analysis.[24]

[24]Obviously, I have no way of knowing whether the order of Erikson, Wright, and McIver's (1993) presentation of the regression results actually reflects the sequence in which the various models were estimated. The fact that they reported "significance" levels for incremental improvements in R^2 from one model to the next suggests that they attached some importance to the ordering, at least for expository purposes. The fact that the Elazar culture variables do not appear in earlier analyses of the same data, including one entitled "State Political Culture and Public Opinion" (Erikson, McIver, and Wright 1987; also Erikson, Wright, and McIver 1989; Wright, Erikson, and McIver 1987), suggests that they were relative latecomers to the analysis, in which case Model 4 in Table 3 may belong earlier in the sequence, making the posterior support for the importance of the culture variables given search-resistant model priors even weaker than I suggest in the text.

A comparison of the alternative mixture posteriors summarized in the various columns of Table 4 does indeed suggest that the model averaging results are quite insensitive to the precise form of the model priors. The posterior point estimates for the baseline effect of state opinion are all within 10% of the estimate produced by uniform model priors, while all of the remaining coefficients (other than the intercept) continue to fall far short of ''statistical significance,'' with average t-statistics of 0.5, 0.4, and 0.4, respectively, from the slightly, moderately, and strongly search-resistant priors.

The implications of these results for inferences about the impact of public opinion within each of Elazar's three types of state political culture are illustrated in Figure 3.[25] Panel (a) shows the posterior distributions for the effects of public opinion in moralistic, traditionalistic, and individualistic states, respectively, implied by Erikson, Wright, and McIver's ''best'' model. The distinction between individualistic states, where public opinion has a very strong effect, and moralistic and traditionalistic states, where the effect is clear but much weaker, is evident.

Panel (b) in Figure 3 shows the corresponding mixture posterior distributions produced by uniform priors over the six models presented in Table 3. Here the apparent effect of public opinion in individualistic states is a good deal weaker, and the apparent effect in moralistic and traditionalistic states is noticeably stronger, producing a virtual overlap of the three distinct distributions. Panel (c) shows the corresponding mixture posterior distributions produced by strongly search-resistant priors over the same six models. The distributions are virtually identical to those in panel (b), except that all of the standard deviations are slightly smaller.

A figure illustrating alternative posterior distributions for the baseline effects of Elazar's three state cultures would show a similar pattern, with distinctly more liberal policy outcomes in individualistic states in the posteriors derived from Erikson, Wright, and McIver's preferred model, but virtual overlap among the distinct posterior distributions produced by any of

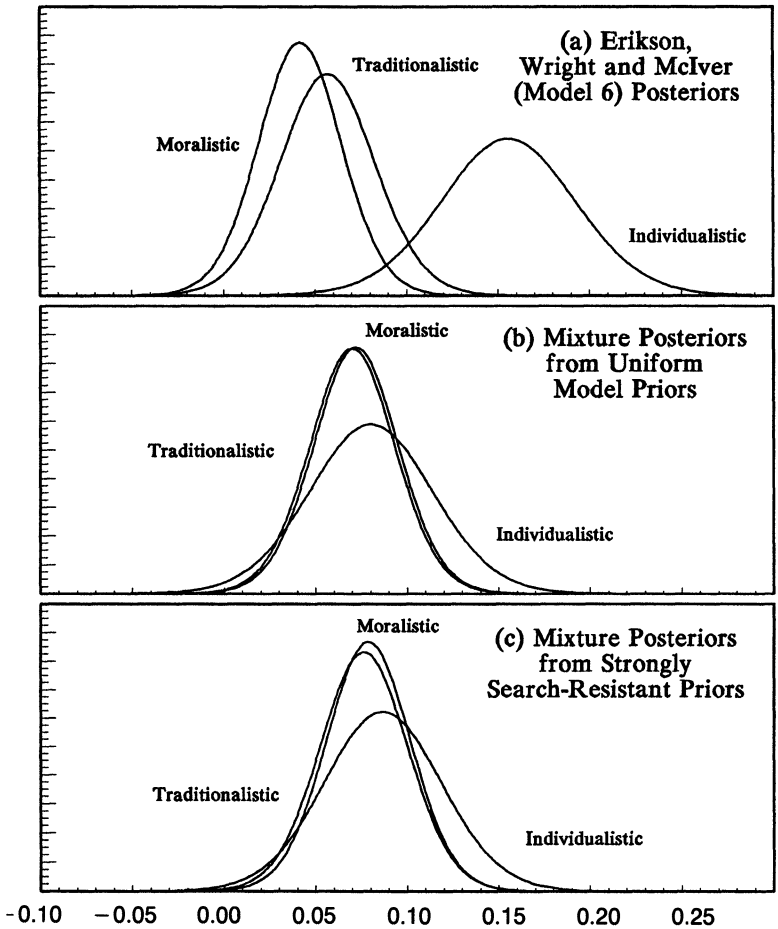
[25]The conditional effects of public opinion displayed in Figure 3 are derived from the means, variances, and covariances of the mixture posterior distributions summarized in Table 4 using the standard formula

$$V(\mathbf{b}_i + \mathbf{b}_j) = V(\mathbf{b}_i) + V(\mathbf{b}_j) + 2\,\mathrm{Cov}(\mathbf{b}_i, \mathbf{b}_j).$$

By an obvious generalization of Equation [6], the covariances of the mixture posterior distributions are the off-diagonal elements of the matrix

$$V(\boldsymbol{\beta}\,|\,\mathbf{X}, \mathbf{y}) = \sum_j \pi_j\, V(\mathbf{b}_j) + \sum_j \pi_j\, [\mathbf{b}_j - \bar{\mathbf{b}}][\mathbf{b}_j - \bar{\mathbf{b}}]'.$$

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



the various model-averaging priors. A figure illustrating alternative posterior distributions for the conditional effects of Democratic Legislative Strength would show a noticeably stronger impact in moralistic states in the posteriors derived from Erikson, Wright, and McIver's preferred model, but, once again, virtual overlap among the distinct posterior distributions produced by any of the various model-averaging priors. Thus, the "strong support—sometimes startlingly strong support—for Elazar's formulation"

(Erikson, Wright, and McIver 1993, 175) evident in the Traditionalism and Moralism coefficients and associated interaction terms in Models 2, 3, 5, and 6 completely evaporates when due account is taken of the specification uncertainty inherent in an analysis as complex as Erikson, Wright, and McIver's.

On the other hand, Erikson, Wright, and McIver's main result—that state policies in general are quite sensitive to variations in public opinion—handsomely survives the rather stringent test imposed by the variety of model-averaging analyses presented in Table 4. The posterior point estimates for the baseline effect of state opinion implied by the search-resistant model priors are all of about the same magnitude as with uniform model priors, but slightly more precise. In effect, then, the results of the model-averaging analyses suggest that the various alternative specifications explored by Erikson, Wright, and McIver fail to provide convincing support for Elazar's classification of state political cultures, but do provide impressive evidence of the robustness of the basic relationship between public opinion and state policy estimated in Model 1 in Table 3.

In that case, why not simply retreat to the original Model 1 in Table 3, with state policy liberalism regressed on state opinion and an intercept? The apparent result would be a bigger estimated effect of state opinion (.1137 versus .0792 for the mixture posterior based on uniform model priors) with a substantially smaller standard error (.0114 versus .0336). If our only aim is to report "significant effects," there is little to argue against such a strategy. But if we aspire to produce credible statistical inferences, we need standard errors that reflect our real uncertainty about the magnitudes of the relevant effects, including uncertainty deriving from ambiguities of model specification as well as from stochastic variation in the observed data. In that case, any of the various results presented in Table 4 will very likely be preferable to the original Model 1—or the *a posteriori* most likely Models 2 or 4, or Erikson, Wright, and McIver's "best" Model 6, or perhaps any single model—taken alone.

Discussion

In each of the examples considered here, model averaging seems to shed some useful light on the inferential implications of specification uncertainty in a research setting with very limited data and weak but suggestive theory. Given the prevalence of such research settings in many areas of social science, the potential scope for fruitful application of model-averaging procedures seems wide indeed. Having said that, it is important to recognize that the examples considered here also illustrate some important limitations of systematic model-averaging procedures. The assignment of positive prior probability to five or six models that someone considered interesting enough to estimate and report falls far short of capturing all the complexities of real specification searches.

For example, the "search-resistant model priors" illustrated in Figure 2 take no account of the susceptibility of broad avenues of specification search to be shut off by discouraging results at an early stage of a sequential analysis. Table 1 provides one example of this phenomenon. Lange and Garrett estimated one model with Left Party Vote as a measure of the political strength of labor and another, otherwise identical, model with Left Party Cabinet Portfolios as an alternative measure of the same theoretical variable. Since the second model fit the data much better than the first, the Left Party Vote variable was dropped, and all of Lange and Garrett's (and Jackman's) subsequent models employed Left Party Cabinet Portfolios as the relevant measure of the political strength of labor.

If Model 1 was as plausible as Model 2 *a priori,* presumably a specification similar to Model 3 but with Left Party Vote substituted for Left Party Cabinet Portfolios should also have been as plausible as Model 3 *a priori.* But since that specification was (apparently) never estimated, its prior probability is implicitly set to zero in the model-averaging calculations. A more realistic representation of the range of *a priori* plausible models might include a large number of these "phantom" models reflecting abandoned avenues of alternative model specification, and the (presumably disappointing) results from these additional models would further dilute the inferential weight of the "best" model or models. (Of course, if none of these models fits the data well, there would have to be a *very* large number of them to have much impact on the results, since models with small posterior probabilities get little weight in the model-averaging calculations.)

This is one instance of a more general problem in the implementation of model-averaging procedures: the assignment of positive prior probabilities to a discrete, usually small, set of distinct alternative models can provide only a rough reflection of real specification uncertainty. Of course, a rough reflection is very likely better than none at all, and we would be foolish to make the best the enemy of the good. It seems clear that Erikson, Wright, and McIver's presentation of six different sets of regression results provides a good deal more information than a presentation of any one set of results could have done. Given the prevalence of specification uncertainty in social scientific work, data analysts should strive to explore and convey through sensitivity analysis the implications of a range of plausible models.

The fact that informal sensitivity analysis can go astray, even in the hands of very sophisticated data analysts, provides one strong argument for more systematic procedures of the sort described here. It should not be taken as an argument for avoiding sensitivity analyses, or for pretending to avoid them by presenting the results of a single model specification as if it was the only plausible specification or the only one examined. Nevertheless, the incompleteness of the set of alternative models actually likely to be encompassed in a typical application of model-averaging techniques is an important limitation in principle, and may also be an important limitation in practice.

On the other hand, a substantial winnowing of the set of potentially relevant models on substantive grounds seems likely to reflect a good data analyst's real specification uncertainty much better than any mechanical averaging over all the models that could logically be constructed from a given list of potential explanatory variables. Whereas an agnostic statistician analyzing the state opinion and policy data presented above might assign equal prior probabilities to all 1,023 possible models defined by distinct combinations of the 10 explanatory variables in Table 3,[26] it seems to me that Erikson, Wright, and McIver's selection of six of those 1,023 possible models embodies a great deal of valuable substantive insight—precisely the sort of complex, hard-to-quantify, but quite legitimate auxiliary information that Bayesian analysis should strive to incorporate.

Of course, it is possible to question whether the assignment of prior probabilities to discrete models is the most fruitful way to incorporate data analysts' auxiliary information (and residual uncertainty) about the substance of their problems. A more direct Bayesian approach would be to begin with a model sufficiently elaborate to capture everything of potential interest in the data, and employ proper prior distributions to express *a priori* beliefs about the relative plausibility of various effects.[27] Rather than including a prospective explanatory variable in some models but omitting it from others to express skepticism about its relevance or importance, an analyst would include all potentially relevant variables, functional forms, and so on in a single model but choose a prior distribution that concentrated probability in parts of the parameter space where the values of many parameters are near zero.

[26]For example, Raftery's publicly available software for regression model averaging automatically averages over all of the 2^K models that can logically be derived from a given set of K potential explanatory variables, assigning each of these models equal prior probability.

[27]A thoroughgoing Bayesian would never knowingly choose to attach zero prior probability to a potentially interesting parameter value or hypothesis, since any parameter value or hypothesis with zero prior probability must get zero posterior probability, in which case there is no hope of learning from data. As Draper (1995, 55) put it, "The main way to avoid noticing after the fact that a set of modelling assumptions, different from those originally assumed, turned out to be correct is for one's model prospectively to have been sufficiently large to encompass the retrospective truth." Unfortunately, a model sufficiently large to encompass the "truth" with certainty will be too large to be manageable (or estimable).

This full-blown Bayesian approach has the theoretical and practical advantage of avoiding unrealistic spikes of posterior probability attached to specific parameter values (as in Figure 1), since it does not assign positive prior probability to models that are dominated by more general models in which they are nested. Unfortunately, it is virtually impossible to implement the full-blown Bayesian approach in practice, since the difficulty of specifying a realistic proper prior distribution is formidable even for relatively simple problems, and no problem remains simple once we commit ourselves to considering "all potentially relevant variables, functional forms, and so on."

If the model-averaging approach makes sense, it must be as a convenient approximation to this more elegant but unrealistic strategy of full-blown Bayesian analysis of suitably elaborate models. In effect, the model-averaging approach represents a further evolution of Leamer's (1978, 15) efforts to patch together a reconciliation of statistical theory and data analytic practice:

> There is no doubt in my mind that uncertain prior information is used to analyze nonexperimental data. But there is also no doubt in my mind that uncertain prior information is impossible to quantify precisely. Ad hoc procedures may, in fact, be efficient methods of using imprecisely defined priors.

Much additional experience will be required to determine whether the procedures described here are, indeed, "efficient methods of using imprecisely defined priors." In the meantime, however, they clearly have three important points in their favor. First, they seem a good deal less *ad hoc* than the murky specification searches and intuitive syntheses described by Leamer, and still practiced even (especially?) by sophisticated data analysts. Second, they are surprisingly simple to implement: all of the calculations in this paper could have been (and most were) produced from standard regression output and a pocket calculator. Finally, they take as their primary raw material what real data analysts are already producing in prodigious quantities—alternative model specifications. Given the evident failure of more conventional attempts to make sense of that raw material, even an imperfect method of calculating the implications of specification uncertainty may be capable of shedding a good deal of light on the fundamental question of what we can learn from our data.

Manuscript submitted 15 August 1995.
Final manuscript received 1 June 1996.

APPENDIX

This Appendix provides the data on which the regression analyses presented in Tables 1 and 3 are based. Table A1 provides the data for Lange and Garrett's (1985, 1987) and Jackman's (1987) analyses of politics and economic growth in 15 OECD countries in the late 1970s. Table A2 provides the data for Erikson, Wright, and McIver's (1993) analysis of public opinion and policy outcomes in the American states. Readers are referred to these sources for more detailed descriptions of the data.

Table A1. Data on Economic Growth in OECD

	Change in Economic Growth	Labor Organization Index	Left Party Vote Share	Left Party Cabinet Portfolios	Dependence on Imported Oil
Australia	.51	1.87	.447	.305	.142
Austria	.64	3.06	.467	1.000	.395
Belgium	.44	2.80	.329	.210	.594
Canada	.50	.98	.152	0	.032
Denmark	.36	2.77	.316	.755	.834
Finland	.56	2.76	.432	.402	.563
France	.57	.68	.406	.017	.628
West Germany	.53	1.80	.405	.748	.510
Italy	.53	1.47	.395	.065	.704
Japan	.38	.43	.314	0	.751
Netherlands	.44	1.90	.313	.412	.579
Norway	1.05	3.33	.474	1.000	−.241
Sweden	.44	3.52	.505	.459	.577
United Kingdom	.26	1.81	.429	.860	.271
United States	.51	.82	0	0	.208

Change in Economic Growth: average percentage change in GDP, 1974–80, divided by average percentage change in GDP, 1960–73 (Lange and Garrett 1985, Table 1). **Labor Organization Index:** additive index derived from labor force unionization and centralization (Lange and Garret 1985, Table 2). **Left Party Vote Share:** average proportion of total vote gained by left parties, 1960–80 (Lange and Garrett 1985, Table 2). **Left Party Cabinet Portfolios:** average proportion of cabinet portfolios held by left parties, 1974–80 (Lange and Garrett 1985, Table 2). **Dependence on Imported Oil:** net oil imports divided by total energy requirements, 1974–80 (computed from *Energy Balances of the OECD Countries, 1971–1981,* Paris: OECD, 1983).

Table A2. Data on State Policy Liberalism

	Policy Liberalism	State Opinion	Legislative Liberalism	Dem Leg Strength	Elazar Culture
AL	−1.45	−23.1	−.623	.968	Traditionalistic
AZ	−1.05	−18.2	−.177	.396	Traditionalistic
AR	−1.54	18.3	.645	.951	Traditionalistic
CA	1.49	−6.3	3.440	.631	Moralistic
CO	.48	−8.6	.036	.408	Moralistic
CT	1.19	−4.4	2.940	.626	Individualistic
DE	1.11	−12.2	.070	.566	Individualistic
FL	−.37	−17.1	−.088	.733	Traditionalistic
GA	−1.04	−17.7	−.458	.890	Traditionalistic
ID	−.87	−27.9	−2.930	.338	Moralistic
IL	.41	−10.1	.943	.537	Individualistic
IN	−1.20	−16.7	−.478	.421	Individualistic
IA	.44	−13.5	2.001	.499	Moralistic
KS	.24	−15.9	.649	.466	Moralistic
KY	−.32	−13.2	.217	.759	Traditionalistic
LA	−1.04	−23.0	−.821	.953	Traditionalistic
ME	−.02	−14.7	1.583	.528	Moralistic
MD	.85	−5.7	2.956	.763	Individualistic
MA	1.64	−.8	4.174	.792	Individualistic
MI	1.18	−8.8	2.372	.525	Moralistic
MN	.79	−12.8	3.159	.632	Moralistic
MS	−1.51	−25.4	−.943	.952	Traditionalistic
MO	−.55	−15.5	−.382	.675	Individualistic
MT	.60	−11.1	−.354	.502	Moralistic
NH	−.14	−12.8	.482	.431	Moralistic
NJ	1.34	−3.4	2.690	.598	Individualistic
NM	−.99	−16.0	−.825	.642	Traditionalistic
NY	2.12	−3.1	2.126	.504	Individualistic
NC	−.96	−20.7	.138	.873	Traditionalistic
ND	−.52	−26.6	−.607	.358	Moralistic
OH	.64	−10.1	.486	.494	Individualistic
OK	−.98	−27.3	−1.441	.760	Traditionalistic
OR	1.39	−7.9	3.207	.667	Moralistic
PA	1.01	−10.6	1.469	.520	Individualistic
RI	.68	−2.1	2.212	.825	Individualistic
SC	−1.53	−21.4	.788	.888	Traditionalistic
SD	−.95	−24.1	−1.990	.295	Moralistic
TN	−.85	−16.6	−.745	.564	Traditionalistic
TX	−.65	−23.2	.469	.820	Traditionalistic
UT	−.44	−28.0	−2.556	.323	Moralistic
VT	.79	−11.4	1.537	.421	Moralistic
VA	−.84	−17.9	−.482	.663	Traditionalistic
WA	.35	−5.9	1.381	.542	Moralistic

Table A2. *(continued)*

	Policy Liberalism	State Opinion	Legislative Liberalism	Dem Leg Strength	Elazar Culture
WV	.12	−9.2	.931	.824	Traditionalistic
WI	1.23	−10.5	3.102	.617	Moralistic
WY	−.70	−17.8	−1.737	.383	Individualistic

Policy Liberalism: additive index derived from standardized scores on eight policy measures (education spending, Medicaid eligibility, AFDC eligibility, consumer protection legislation, criminal justice legislation, legalized gambling, ERA ratification, and tax progressivity), circa 1980 (Erikson, Wright and McIver 1993, 77, Table 4.2). **State Opinion:** percent liberal ideological identification minus percent conservative identification, accumulated from 122 CBS News/*New York Times* polls, 1976–88 (Erikson, Wright, and McIver 1993, 16, Table 2.2). **Legislative Liberalism:** "weighted average of Democratic and Republican elite ideology scores [Erikson, Wright, and McIver 1993, 103, Table 5.3], where the weights are determined by the parties' relative legislative strength [as measured by Democratic Legislative Strength]" (Erikson, Wright, and McIver 1993, 128). **Democratic Legislative Strength:** average proportion Democratic of state legislature (weighting upper and lower houses equally), 1977–84 (*Statistical Abstract of the United States, 1985*, Washington: U.S. Bureau of the Census, 1984). **Elazar Culture:** Elazar's (1972) classification of state political cultures as "individualistic," "moralistic," or "traditionalistic" (Erikson, Wright, and McIver 1993, 172, Figure 7.4).
Following Erikson, Wright, and McIver (1993), Alaska, Hawaii, and Nebraska are omitted due to missing data and Nevada is omitted due to an implausibly liberal State Opinion score.

REFERENCES

Achen, Christopher H. 1982. *Interpreting and Using Regression.* Beverly Hills: Sage Publications.

Adams, J. L. 1991. "A Computer Experiment to Evaluate Regression Strategies." *Proceedings of the Computer Statistics Section of the American Statistical Association* 55–62.

Alvarez, R. Michael, Geoffrey Garrett, and Peter Lange. 1991. "Government Partisanship, Labor Organization and Macroeconomic Performance." *American Political Science Review* 85:539–56.

Bartels, Larry M. 1990. "Five Approaches to Model Specification." *The Political Methodologist* 3:2–6.

Beck, Nathaniel, and Jonathan Katz. 1992. "Model Assessment via Cross Validation." Presented at the 1992 annual meeting of the Political Methodology Society, Cambridge, MA.

Beck, Nathaniel, Jonathan N. Katz, R. Michael Alvarez, Geoffrey Garrett, and Peter Lange. 1993. "Government Partisanship, Labor Organization and Macroeconomic Performance: A Corrigendum." *American Political Science Review* 87:945–8.

Draper, David. 1995. "Assessment and Propagation of Uncertainty" (with discussion). *Journal of the Royal Statistical Society, Series B* 57:45–97.

Elazar, Daniel. 1972. *American Federalism: A View from the States.* 2nd ed. New York: Crowell.

Erikson, Robert S., John P. McIver, and Gerald C. Wright, Jr. 1987. "State Political Culture and Public Opinion." *American Political Science Review* 81:797–813.

Erikson, Robert S., Gerald C. Wright, Jr., and John P. McIver. 1989. "Political Parties, Public Opinion, and State Policy in the United States." *American Political Science Review* 83:729–50.

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

Freedman, David A. 1983. "A Note on Screening Regression Equations." *American Statistician* 37:152–5.

Green, Donald P. 1990. "On the Value of Not Teaching Students to Be Dangerous." *The Political Methodologist* 3:7–9.

Jackman, Robert W. 1987. "The Politics of Economic Growth in the Industrial Democracies, 1974–80: Leftist Strength or North Sea Oil?" *Journal of Politics* 49:242–56.

Jeffreys, Harold. 1961. *Theory of Probability.* 3rd ed. Oxford: Clarendon Press.

Judge, George G., and M. E. Bock. 1978. *The Statistical Implications of Pre-Test and Stein Rule Estimators in Econometrics.* New York: North-Holland.

Judge, George G., W. E. Griffiths, R. Carter Hill, Helmut Lütkepohl, and Tsoung-Chao Lee. 1985. *The Theory and Practice of Econometrics.* 2nd ed. New York: John Wiley and Sons.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90:773–95.

Kass, Robert E., and L. Wasserman. 1995. "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association* 90:928–34.

Lange, Peter, and Geoffrey Garrett. 1985. "The Politics of Growth: Strategic Interaction and Economic Performance in the Advanced Industrial Democracies, 1974–1980." *Journal of Politics* 47:792–827.

Lange, Peter, and Geoffrey Garrett. 1987. "The Politics of Growth Reconsidered." *Journal of Politics* 49:257–74.

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

Raftery, Adrian. 1994. "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models." Technical Report No. 255. Department of Statistics, University of Washington.

Raftery, Adrian. 1995. "Bayesian Model Selection in Social Research" (with discussion). In *Sociological Methodology 1995,* ed. Peter V. Marsden. Cambridge, MA: Blackwells.

Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6:461–4.

Western, Bruce. 1996. "Model Uncertainty in Macrosociology." In *Sociological Methodology 1996,* ed. Adrian Raftery. Cambridge, MA: Blackwells.

Western, Bruce, and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88:412–23.

Wright, Gerald C., Jr., Robert S. Erikson, and John P. McIver. 1987. "Public Opinion and Policy Liberalism in the American States." *American Journal of Political Science* 31: 980–1001.