Lessons in Disguise: Multivariate Predictive Mistakes in the study of Repeated Collective Choice

Bruce A. Desmarais
University of North Carolina at Chapel Hill
bdesmara@email.unc.edu
March 24, 2010

Abstract

Many of the central processes studied by political scientists involve repeated collective choices made by stable sets of actors – voting in legislatures, decisions issued by multi-member courts of appeal, states considering conflict intervention, etc. While much has been learned, more can be gained from examining situations in which existing statistical models perform poorly, meaning they provide poor predictions about real data. The inherent multivariate interdependence among the actors in this data is often missed by conventional approaches to analysis. I introduce a new prediction-error concept, the joint prediction error (JPE), which is a grouping of simultaneous individual outcomes that is poorly predicted by a model. Moving beyond just poorly predicted individual events, JPEs capture the intersecting information and inter-relationships missed by traditional diagnostics. I provide a benchmark for identifying joint prediction errors. Analysis of JPEs is shown to improve empirical models from two published articles, one on U.S. Supreme Court voting, and another on the fulfilment of international defense alliances.

Note: Replication code of the substantive applications will be made available with permission from the original author(s). I would like to thank Thomas Carsey, Skyler Cranmer, James Stimson and Jeff Harden for their helpful suggestions with this project. This is a work in progress, so please do not cite without the author’s permission.
“The applied statistician should avoid models that are contradicted by the data in relevant ways - frequency calculations for hypothetical replications can monitor a model’s adequacy and help to suggest more appropriate models.”

-Donald B. Rubin (1984)

1 Repeated Collective Decisions

Every field of empirical political science accumulates knowledge through the analysis of data that arise in the form of repeated collective decisions. Roll-call votes in legislatures and decisions issued by multi-member courts of appeal are stable groups of political actors issuing individual decisions that are aggregated into salient collective outcomes. In the international arena, intervention into civil wars, the provision of relief for natural disasters, and the issuance of trade sanctions are interdependent decisions rendered repeatedly by a stable group of states. Due to the importance of the collective outcomes that result from individual decisions (e.g. laws written or the results of civil wars), and the fact that actors have multiple opportunities to learn optimal strategies for interaction, patterns of dependence or relationships are likely to emerge in repeated collective choice data. However, in political science applications we often pool the members of the stable group into a sample for regression modeling where relationships between the members are ignored to the extent that they are not represented by independent variables. If such relationships do exist (e.g. if members of the U.S. House of Representatives learn over time to take cues from certain policy specialists or party leaders, and issue roll-call votes in the same direction as these leaders),

\[^{1}\text{Many scholars have noted that patterns of sophisticated rational interaction are likely to emerge when collective choice situations are repeated many times, and actors can learn the rules and payoffs of the game (see e.g. Verba (1961) and Ostrom (1998)).}\]
statistical inferences from pooled regression are subject to misspecification bias. Since the data contain repeated observations of collective behavior, it can be used to learn about interdependence among the actors. I propose an iterative method for learning and modeling these dependencies. Similar in structure to the approach advocated by Achen (2005), rather than estimate an overly complicated model from stage one, I suggest specifying a simple model to start, and updating it to address predictive deficiencies, subjecting the updated model to rigorous conservative tests of the validity of the innovations.

In many instances of repeated collective decision making, the salient collective outcome (e.g. the result of a court case or the passage of legislation) is a deterministic function of the individual decisions rendered by the members of the collective (e.g. a lower court decision is reversed by the U.S. Supreme Court if five or more justices vote with the appellant). Because of this deterministic relationship between micro and macro-level outcomes, working within either the likelihood or Bayesian estimation frameworks, or any other method used to fit a full parametric distribution to the data, if a model is fit to the individual decisions in the collective, one is automatically implied for the collective outcome. For instance, if a model is fit to U.S. Supreme Court Justice voting on the merits – estimating the probability that a justice will vote with the appellant in a case – the probability that the Court decides in favor of the appellant is given by the sum of the joint probabilities of all nine-justice configurations in which at least five vote with the appellant.

Due to the deterministic relationship between the micro and macro-level models of collective decisions, from the standpoint of probability theory, it is inconsistent to specify separate statistical models for individual decisions and higher-order configurations of those decisions – the the former implies the latter. The critical implication of the micro-macro connection is that, in order to be correctly specified, the micro-level model must capture any tendency for individual decisions to produce sophisticated/intentional higher-order configurations. For instance, Hix, Noury and Roland (2005) finds that there are varying levels of political party
cohesion in the European Parliament. If the findings of Hix, Noury and Roland (2005) are valid, any micro-level analysis of roll-call voting in the European Parliament is misspecified if it does not account for a varying tendency towards intra-party cohesion in members’ votes. Extending an individual-level decision model – often logistic regression where observations are assumed to be independent conditional on the covariates – to allow for flexible forms of interdependence commonly requires non-trivial and at times prohibitive computational effort to estimate the dependence parameters (see e.g. Alvarez and Nagler (1998) for an example where preferences for electoral candidates are posited to be correlated, Franzese and Hays (2007) for a discussion of the estimation challenges in accounting for spatial dependence in time-series cross-section data, and Ward, Siverson and Cao (2007) who find that latent reciprocal and transitive tendencies characterize international dyadic data). Rather than attempt to extend a micro-level model to accommodate every configurational tendency that has either found support in the literature or can be reasonably conceived, I develop a method to identify configurations that are missed by a simple micro-level model – the analysis of which suggests the key configurational extensions necessary to make valid inference on the micro-level processes. The key mechanism underlying this procedure is a method of residual analysis that is particularly suited for repeated collective choice data.

The systematic analysis of residuals from regression models has long been used to monitor aspects of statistical fit with the goal of improving specification (see e.g. Cox and Snell (1971), Achen (1977), Beck (1982) and Achen (2005)). Generally speaking, residual analysis involves the comparison of observed data with predicted values. There are at least two major general challenges in residual analysis. First, with the generic goal of assessing the proximity of observed and predicted quantities, the particular function(s) of the observed and predicted data to be compared (e.g. expected value, variance, correlation, skewness etc.) must be insightfully chosen. Second, the analyst must specify the level of divergence between the true and predicted quantity that constitutes an interesting deficiency. I introduce the
concept of a joint prediction error (JPE), which is a collective outcome that is observed to occur with a much different frequency than predicted, and provide benchmarks for deciding what constitutes a JPE. In doing so, I overcome a particular challenge that arises in the analysis of these joint residuals. Given \( n \) members of the collective, and interest in finding JPEs composed of \( k \) members, \( \binom{n}{k} \) groups need to be considered, which can become a giant number for realistic size collectives and even small \( k \). For instance, if one is interested in monitoring predictive accuracy of a model predicting legislative activity on all groupings of five legislators in a 435 member chamber, there are 126,837,202,212 groups to check, and if \( k \) is increased to ten the number of groupings is multiplied by 474,925,189. I introduce algorithms from the machine-learning literature, designed to find frequent joint occurrences in databases of millions of commercial transactions, and show that they can be used to efficiently search over all possible combinations of actors in the collective.

Finally, by replicating and extending two recently published studies, I demonstrate how improvements in discrete choice models of repeated collective choice processes can be discovered through the analysis of JPEs. I find that a logit model explaining Supreme Court votes on the merits published by Johnson, Wahlbeck and Spriggs (2006) critically understates the degree of case-level consensus on the Court. This observation leads to an improved model specification that accounts for correlation between the justices and includes additional important case-level covariates. In Gartzke and Gleditsch (2004), a study on international defense alliance activation, similar to the inter-justice correlation in the Supreme Court example, the empirical model understates association among a state’s allies. Additionally in the defense alliance application, a pattern emerges in the JPEs which suggests that states with greater consultation obligations are less likely to enter a war in defense of their allies. Adding a measure of a state’s consultation obligations to the model in Gartzke and Gleditsch (2004) (1) supports the insight that states with more consultation pacts are less likely to support their allies and (2) suggests that the original central empirical finding of the article
– that democratic states are less likely to assist their allies – resulted from the omission of consultation obligation. In both replications, the published statistical analyses are improved by extensions suggested by the JPEs. Improvement is verified through multiple model fit metrics.

2 Information, Data and Repeated Collective Choice

As noted previously, most collective choice modeling in political science involves an intense focus on the numerous interactions of a relatively small set of actors. This form of scholarship creates two conditions that catalyze theoretical innovations from the simple visual inspection of prediction errors. The first consequence is that simple labels – country names, legislative districts, justice names, etc.– on the actors in the dataset communicate information to the analyst above and beyond that which is contained in the rows and columns of the dataset. The second is that there is quite likely to be an overwhelming amount of previous theoretical and empirical research that precedes any new study of historical political data. Both of these features present unique opportunities for improvement with joint prediction error analysis.

In their analysis of the representational efficacy of majority-minority Congressional districts, (Cameron, Epstein and O’Halloran 1996, pp. 810) state, “In many southern state legislatures, [minority group leaders and Republicans] formed voting blocs when passing redistricting plans, and the [U.S.] Justice Department under Republican presidents was eager to create the maximum possible number of majority-minority districts.” This represents rich information about the process under study – the motivations underlying the formation of majority-minority districts – yet no data or citation to outside work is provided. It is knowledge held by the authors, the validity of which was accepted at face-value by reviewers at the American Political Science Review (APSR). Anyone who has presented at a conference, and
been confronted with the one case (e.g. legislator, country, year) that represents the perfect counter-factual to his or her theory, knows that political scientists have auxiliary expertise – constituting information about the observations above and beyond that which appears in the regression equation. If a scholar of civil war intervention runs a logistic regression model on the intervention decisions of states, he or she may recognize that the model poorly predicts outcomes in which developed states decide to intervene and others do not without collecting additional data about countries. Such a recognition would serve as motivation to collect and include in the model a measure of a state’s development. This auxiliary information optimizes potential benefits from simply examining those combinations that are poorly predicted by a given statistical model.

The second consequence of multiple studies of familiar observations is that the discipline accumulates a predictable set of control variables that are considered potentially serious omissions if left out of a model. For most salient topics in political science, dozens of studies precede any new research. Most of these studies propose partially unique explanations of a process and thus provide candidate control variables for anyone who endeavors to model the same or similar data in the future. It is uncommon and practically infeasible for one to include every variable that has ever been found to significantly influence a process in a new analysis. At the same time, previous findings cannot be ignored simply for the sake of time or parsimony. Examining joint prediction errors constitutes a reasonable compromise between ignoring past work outright and including the entire preceding empirical literature in an initial model. Knowledge about the approximate values of the omitted factors can be checked for consistency with patterns in the JPEs. For instance, judicial scholars are familiar with the seniority ranking of justices on the U.S. Supreme Court. Analysis of joint errors from a model of Supreme Court voting would reveal whether justices close in seniority were voting similarly.
3 Iterative Model Improvement Through Prediction

Error Analysis

The process I prescribe for developing the best statistical model of repeated collective choice data is premised by the observation of Rubin (1984), that frequency calculations performed on the real data should not differ from model predictions in relevant ways. Once a model has been fit to data and is treated as the best possible model, it assumes the position of the analyst’s null or assumed model. Quantities in the observed data that differ considerably from predictions drawn from the model serve as evidence against the null. Intuitively, if it is claimed, through the parametric fitting of a model, that a complete distribution for the data is provided, then it should not be possible to find distributional qualities that of the data that contradict the model.

There are essentially five stages in one iteration of the model-fitting procedure I advocate:

1. Fit the model (M) that represents the best specification the analyst can currently manage.

2. Draw many hypothetical datasets according to the probability distribution of the data implied by the model.\(^2\)

3. Identify joint prediction errors by finding combinations of outcomes that occur with much greater or lesser frequency in the data than is predicted by the model.

4. Update the model to accommodate deficiencies that are hypothesized to produce the prediction errors.

\(^2\)It is possible that in simple cases the analytic distribution of the data will be available, but to assure the algorithm is applicable when it is not available, I advocate simulation.
5. Assess, using model-fit metrics that favor a parsimonious specification, whether the updated specification provides a better fit to the data than the previous one.

This process can be repeated indefinitely, or until the analyst has no more intuition about the deficiencies creating the prediction errors. For those wary of purely data-driven procedures for model construction, it is important to recognize the importance of theory in the fourth step. Without a thorough theoretical understanding of the process under study, it will not be possible to recognize the significance of the JPE membership. For instance, a Congress scholar may recognize – through inspection of the JPE memberships – that a model of roll-call voting in the U.S. House poorly explains votes in which members on the Appropriations Committee disagree with those on the Budget Committee. Without at least a loose recollection of committee membership in the House, it would not be possible to even recognize never-mind explain such a pattern. Of course, any data-driven model-fitting procedure must guard against over-fitting the sample data. After presenting the algorithm used to identify JPEs, I present a model fit metric that can be used to avoid over-fitting.

The specific metric used to determine whether a joint outcome constitutes a JPE is the posterior predictive p-value introduced by Meng (1994). Some basic concepts need to be clarified before I can define a posterior predictive p-value. In a Bayesian analysis, the prior distribution of the parameters ($\pi(\theta)$) represents the analyst’s belief about the parameters prior to using the observed data ($X$). The posterior distribution of the parameters ($p(\theta|X)$) is the resulting belief regarding the distribution of the parameters after updating the prior distribution with the observed data. In a Bayesian analysis, point estimates are equal to the means of the posterior distribution, and credible intervals – the Bayesian analog to the frequentist confidence interval – are derived from the quantiles of the posterior distribution (Gill 2002). The posterior distribution conditional on the observed data $X$ is given by
\[ p(\theta|X) = \frac{l(X|\theta)\pi(\theta)}{\int_\Theta l(X|\theta)\pi(\theta)d\theta}, \]

where \( l(X|\theta) \) is the likelihood function of the data given \( \theta \). If \( M \) is fit by maximum likelihood, the asymptotic sampling distribution of \( \theta \) is used as an approximation of the posterior distribution of \( \theta \) (King, Tomz and Wittenberg 2000; Tomz, Wittenberg and King 2003), which is multivariate normal with mean vector equal to the parameter estimates \( (\hat{\theta}) \) and covariance matrix equal to the variance-covariance matrix of \( \hat{\theta} \) (\( \hat{\Sigma} \)). The posterior predictive distribution (PPD) of \( X \) is the expected distribution of future replicates of \( X \). It represents the analyst’s belief about the distribution of \( X \) after updating with the available data. The posterior predictive distribution \( f(X_{new}) \) of the data is computed by averaging the likelihood function over \( p(\cdot|\theta) \), and is given by

\[ f(X_{new}) = \int \_{\Theta} l(X_{new}|\theta)p(\theta|X)d\theta. \]  

In practice, \( p(\theta|X) \) and/or \( f_X(\cdot) \) are often not available in closed form due to intractability of the integrals in equations 1 and 2. In the typical Bayesian analysis, using Markov Chain Monte Carlo (MCMC) methods, the researcher has a large sample from \( p(\theta|X) \) rather than a formula for the posterior distribution. The algorithm given in figure 1 can be used to draw from the posterior predictive distribution using the MCMC sample. When \( M \) is fit by maximum likelihood, the sample from the posterior distribution derived through MCMC is replaced with a random sample from the asymptotic sampling distribution.

P-values are commonly used in political science to measure the plausibility of some null parameter (e.g. population mean, difference in means of two populations, the regression coefficient in the population, the variance etc.) given an observed sample counterpart of
that parameter (i.e. statistic) and additional assumptions about the data-generating process. Suppose it is of interest to assess the oddity or rarity of the observed value of some statistic computed on the data \((T(X))\) given an assumption about the distribution that generates \(X (f(X))\). If it is possible to derive the distribution of \(T(X)\) given \(f(X) (g(T(X)))\) (i.e. the sampling distribution in a classical context), the placement of \(T(X)\) on \(g(T(X))\) can be used to estimate the area under \(g(T(X))\) to the right (left) of a comparatively high (low) value of \(T(X)\) to derive a p-value.

A considerable challenge in many settings is that the analytic sampling distribution (i.e. \(g(T(X)))\) is not available in closed form for many combinations of \(T(X)\) and \(f(X)\). For instance, the analytic sampling distribution of the sample median is rarely available in closed form (Greene 2008, pp. 597). Originally suggested by Rubin (1984) and thoroughly explored by Meng (1994), the posterior predictive p-value provides a general solution for determining the rarity of an observed value of \(T(X)\) given a fully parametric specification of \(f(X)\). If \(T(\cdot)\) is computed on many draws of hypothetical data from \(M\) using the posterior predictive distribution, the empirical distribution of \(T(X)\) over the draws of \(X (h(T(X)))\) can be used as a substitute for \(g(T(X))\). As the number of draws of \(X\) from \(M\) approaches infinity, the tail area outside of \(T(X)\) on \(h(T(X))\) approaches a p-value for \(T(X)\) given \(M\) as the null model.

In the context of joint prediction error analysis, let \(T(X)\) be the number of times a multivariate outcome (\(\Gamma\)) occurs in the data. As a hypothetical example, a possible \(T(X)\) is the number of times Barak Obama and Hillary Clinton voted in the same direction on roll-calls in the U.S. Senate. Using \(M\) as the null model, if \(\Gamma\) has a posterior predictive p-value less than a tunable parameter \(\alpha\), it is classified as a joint prediction error.

To find joint prediction errors in a repeated collective choice dataset the p-value for every possible \(\Gamma\) must be computed. As noted earlier the universe of possible \(\Gamma\)s can be
quite large. This poses a computational challenge in counting the frequency of $\Gamma$ in both
the real and simulated data for all $\Gamma$s. Thankfully, this counting problem is very similar
in structure to a challenge that has been considered in the machine learning literature for
decades – counting, in databases of millions of commercial transactions for merchants offering
thousands of products, the number of times product groupings occur in shopping baskets
(e.g. the number of times a T.V. Guide, fishing pole and neck tie are all purchased together
in transactions at Wal-Mart). *Frequent itemset mining* is the general term that encapsulates
work on finding product groupings that meet certain criteria (Wen 2004; Luo and Zhang
2007). Treating the collective choice as the transaction, and the individual decisions made
by the actors as the product occurrences, frequent itemset mining algorithms can be used to
count the joint occurrence of individual decisions within collective choices. I take advantage
of frequent itemset mining algorithms in the implementations of JPE analysis below. ³

There are three parameters that must be set by the user of the algorithm outlined above:
the size of the joint prediction errors ($k$), the number of draws from $f(X_{new})$ to be used to
compute the posterior predictive p-values ($t$), and the level of the p-value ($\alpha$) at which to
classify the joint outcome as a prediction error. As I will demonstrate through application
later, a great deal of information is communicated in pairs. Pairs contain all of the available
information about what outcomes occur together. For this reason, I suggest a default value
of $k = 2$. It may be informative to move beyond $k$ if particular higher order configurations

³Many of the algorithms available in the R package *arules* (Hahsler, Grn and Hornik 2005)
can be combined to efficiently implement JPE analysis in large datasets. I am developing an
R package (*JPEMiner*), in which I wrap and structure a number of the algorithms in *arules* to
efficiently perform JPE analysis after the estimation of many discrete choice models familiar
to political scientists.
are of interest. For instance, if one were interested in assessing whether a model accurately predicted intra-continental agreement in U.N. Security Council votes, it would be possible to look at the pairwise predicted versus observed agreements among all pairs within a continent, but it might be easier to compare the predicted and observed occurrences of continent-level consensus. The term $\alpha$ should be chosen to produce a manageable set of prediction errors – not so low that no prediction errors are discovered, and not so high that every joint outcome is considered a prediction error. Lastly, $t$ should be set high (1,000–10,000) to start, and the analysis should be repeated two or three times to assure the results are not attributable to simulation error. If results differ across repetitions, $t$ should be increase until variation across repetitions is negligible.

The suggestions provided in the previous paragraph represent reasonable starting points for most applications, but should not be read as strict constraints on the values of the tuning parameters. It is important to emphasize that discovering a pattern in the JPE analysis does not constitute rigorous statistical inference on the factors creating that pattern. The validation step comes after $M$ has been updated to account for patterns discovered in the JPE analysis. The objective in the JPE analysis stage is to tune the parameters $(k, t, \alpha)$ until either some intuition is reached regarding appropriate improvements to $M$ or it is clear that no meaningful discrepancy between the data and the distribution implied by $M$ can be found. The point is to push $M$ to the breaking point in regards to its consistency with the data, with the intention of reconstructing a stronger model through a theoretical account of the prediction errors produced by $M$. The validation procedure presented next is used to judge the validity of the proposed updates to $M$. 

13
3.1 Evaluating the Updated Model

As noted previously, observing a pattern in the joint residuals does not constitute statistical confirmation of that pattern as a component of the data generating process. Since the method of model improvement proposed here is fairly data-intensive, it is desirable to use a relatively conservative method of evaluating the fit improvement associated with the updates, so as to avoid over-fitting. The method I advocate is cross-validation.

Cross-validation avoids over-fitting by evaluating the fit of a model on data that was not used to estimate the parameters of the model. The parameters of competing models are estimated on the training set, and the relative fit is judged using the validation set (data that was not used to estimate the model, but is considered to be drawn from the same population as the training set) (Jensen and Cohen 2000). Leave-one-out cross-validation is a method of judging the predictive fit of a generalized model (GLM) where every observation is iteratively used in the training and validation sets, and therefore does not require the analyst to arbitrarily exclude some of the data from estimation (Snee 1977; Burman 1989; Thall, Simon and Grier 1992).

In order to implement cross-validation, a predictive measure of the fit of the model to the excluded observations must be identified. Many candidates have been considered including the cross-validated classification error for categorical outcomes (Leo et al. 1984) and the cross-validated squared error for continuous outcomes (Hjorth 1993). A predictive measure that is particularly useful when the objective is to compare fully parametric models is the cross-validated log-likelihood. The cross-validated log-likelihood (CVLL) is computed by summing the log-likelihood of each observation given the parameters estimated on the rest of the data set \( \theta^{-i} \) (Rust and Schmittlein 1985; O’Sullivan, Yandell and Raynor 1986; Verweij and Van Houwelingen 1993; van Houwelingen et al. 2006). A very common metric of distance between two probability distributions is the Kullback-Leibler distance (Gelman, Meng and
Stern 1996; Clarke 2003, 2007). In expectation, among a number of candidate models, the model with the highest CVLL is that with the minimum Kullback-Leibler distance from the true model (Cover and Thomas 1991; Smyth 2009). Thus, if the updates to $M$ move the specification closer to the true model, then, on average, evaluation with the CVLL will indicate that the updates should be accepted. The formula for the CVLL is given by

$$CVLL = \sum_{i=1}^{N} \ln \left[ l(x_i|\theta^{-i}) \right].$$

(3)

The CVLL is extended to data that is organized hierarchically by clustering on a single level (e.g. court case) – a structure common in repeated collective choice data – by leaving out one cluster at a time and summing the log-likelihood of the left-out clusters rather than leaving out a single observation (Price, Nero and Gelman 1996). To evaluate the fit of the various models specified in the current analysis, I compute the CVLL as well as the BIC, another conservative measure of model fit, in each of the applications below.

4 Replications with JPE-Suggested Extensions

4.1 The U.S. Supreme Court and Oral Argument Quality

Johnson, Wahlbeck and Spriggs (2006) test whether the quality of oral argument before the U.S. Supreme Court influences the votes of the justices. Justice Harry Blackmun graded the oral arguments of attorneys on an 8-point grading scale for cases argued before the Supreme Court from the 1970-1994 terms. Johnson, Wahlbeck and Spriggs (2006) specify a logistic regression model of votes (pooled over justices, cases and terms) where the dependent variable is coded 1 if the justice votes to reverse the lower court decision and 0 for affirm. The votes of Justice Blackmun are excluded due to concerns about endogeneity. A number
of other control variables are included. See the original article for their justification.

4.1.1 Case-Level Prediction Errors

The collective choices made by the justices on the U.S. Supreme Court are case decisions. Each case is represented as a combination of justice-votes. On a typical case, there are eight justices (excluding Blackmun) who can each either vote to affirm or reverse, leading to \(2^8 = 256\) possible eight-vote outcomes. The JPE analysis is performed on the full model specified in column 2 of table 3 in Johnson, Wahlbeck and Spriggs (2006). In the analysis I report I used \(t = 1,000\) draws from the posterior predictive distribution of the data, a posterior-predictive p-value of \(\alpha = 0.10\), and a prediction error size of \(k = 2\) justice-votes.\(^4\) Figure 2 gives the four most frequent over-predicted and under-predicted justice-vote pairs in the dataset. An under(over)-prediction is a pair that is predicted to occur less(more) frequently than it actually does. The left and right columns give under and over-predicted pairs respectively. Each panel is a histogram of the number of cases in which the justice-vote pair occurs in the 1,000 datasets drawn from the original model in Johnson, Wahlbeck and Spriggs (2006). The number of cases in which the pair occurs in the actual dataset is located at the solid vertical line in each panel.\(^5\)

\(^4\)I repeated the analysis with three different simulated samples, and there was no variation in the set of prediction errors – leading me to conclude that the \(t = 1,000\) is sufficiently large to avoid simulation error. Also, the substantive inferences I draw from the JPE analysis do not change for \(\alpha\) as small as 0.05, and there is no utility in using a less restrictive p-value. Lastly, I looked at JPEs of size \(k \in \{3, 4, 5\}\), but gathered no additional intuition regarding model improvement from the larger groups.

\(^5\)R package \texttt{Arules} Michael Hahsler and Hornik (2009) was used to perform the frequent itemset mining. I do not replicate the model in column 1 of table 3 in Johnson, Wahlbeck
Examining figure 2 demonstrates a clear pattern in the prediction errors. All of the under-predicted pairs are justices in agreement. All of the over-predicted pairs are justices not in agreement. The results presented in the figure suggest that the original model heavily under-predicts agreement among justices in their votes on the merits. This pattern is confirmed in the larger set of JPEs. A total of 160 JPEs are identified. Among the 91 under-predicted pairs, 83 are pairs of justices voting in the same direction. The remaining 69 JPEs are over-predictions, and 68 of them are justices voting in opposite directions (i.e. one voting to reverse and one to affirm).

What these findings suggest is that the original model misses a strong degree of positive correlation between the votes of justices on any given case. This is an omitted feature of the data generating process that threatens the validity of inferences through misspecification bias (White 1982). Two classes of underlying mechanisms could be contributing to the observed correlation. First, it is possible that overt influence or cooperation occur on the Court. Previous studies have found that the Court tends towards consensus decision-making (Haynie 1992; Epstein, Segal and Spaeth 2001). It could also be that omitted legal factors are producing correlation. If there are legal facts that point every justice (or a large subset thereof) in a particular direction, the omission of these factors from the model would cause the under-prediction of justices voting in a consensus manner. Case-level apolitical factors are just recently gaining acceptance as important predictors of the votes of Supreme Court justices (Spriggs and Hansford 2001, 2002; Johnson, Wahlbeck and Spriggs 2006). The early dominance of the attitudinal model made light of case-level idiosyncrasy (Segal and Cover and Spriggs (2006) because an LR test strongly rejects the hypothesis that the restrictions in the reduced model are valid.
1989; Segal and Spaeth 1996, 2002). A growing number of studies, such as Collins (2004), Johnson, Wahlbeck and Spriggs (2006) and Collins (2007), indicate that factors acting at the case-level other than political-ideological concerns motivate the choices made by the justices. Consensus prediction errors don’t constitute a statistical test for the presence of unobserved association in justices’ votes. The model from Johnson, Wahlbeck and Spriggs (2006) must be improved to both test and account for positive correlation.

4.1.2 Case-Level Determinants of Supreme Court Votes

I extend the model in Johnson, Wahlbeck and Spriggs (2006) in two ways to account for the pattern discovered in the JPE analysis. First, as mentioned previously, omitted case-level covariates could cause the observed association among the justices. Collins (2004, 2007) show that the Court responds to *Amicus Curiae* briefs. Specifically, it is shown that the probability that a particular side wins a case is directly proportional to the number of briefs filed on its behalf and inversely proportional to the number of briefs filed for the other side. Moreover, briefs filed by the U.S. Solicitor General have a larger effect on the Court’s decisions than do those filed by others. I add a series of variables to the model to account for this. The variables *Appellee Amicus*, *Appellant Amicus*, *SG Appellee Amicus* and *Appellant Amicus* are the number of *Amicus Curiae* briefs filed on behalf of the appellee, appellant, appellee by the Solicitor General and appellant by the solicitor general respectively. Following Collins, I expect that briefs filed on behalf of the appellant (appellee) will have a positive (negative) effect on the likelihood a justice votes to reverse. I also add one more case-level control to the model; *Lower Court Conflict*, an indicator of whether the reason for granting *certiorari* is rooted in lower court conflict. Collins (2004) finds that the Court is less likely to reverse a decision that it hears due to lower court conflict.\(^6\) I expect this variable to have a negative

---

effect on the probability a justice votes for reversal.

The degree of consensus demonstrated in the JPE analysis is quite marked. It would be overly optimistic to assume that all of the case-level association would be explained by the covariates I add to the model. I therefore update the model to explicitly estimate the residual association among the justices’ votes. A standard tool for modeling unobserved cluster-wise association in regression models is to include a hierarchical random-effect in the likelihood function (Gelman and Hill 2007). It is assumed that there is a shared disturbance to the linear predictor to for every observation in a cluster. In the model implemented below, the random effect is assumed to be normally distributed with zero mean. It is integrated out of the likelihood function, leaving only a variance term of the random effect to be estimated. The higher the variance, the higher the correlation between the observations in the same cluster (Caffo, An and Rohde 2007). Thus, the second update to the model presented in Johnson, Wahlbeck and Spriggs (2006) is to add a case-level random effect.

[Insert Table 1 here]

The results of the hierarchical logistic regression models are presented in table 5. The model closest to the baseline specification that appeared in Johnson, Wahlbeck and Spriggs (2006) is the Justice-Level specification. Johnson, Wahlbeck and Spriggs (2006) use cluster-robust standard errors (Williams 2000) with the Justice as the clustering variable. In the case of logistic regression, this covariance estimator produces standard error estimates that are biased downward and the estimator itself is inconsistent in the face of unmodeled hetero-

\footnote{R package \texttt{lme4} (Bates and Sarkar 2006) was used to estimate the models in table 5}
geneity (Greene 2008, p. 517; Harden n.d.), so I use an alternative mechanism to account for within-justice correlation. I add a justice-level random effect to this model. This is compared to a model with a case-level random effect.8

The pattern discovered in the joint prediction error analysis led to a specification that greatly improves model fit, and alters many of the inferences derived from the original model. Adding the case-level random effect to the original model reduces both the CCVLL and BIC by almost 25%. Also there is much more unobserved heterogeneity and/or correlation at the case-level than the justice-level. The case-level random effect variance is estimated to be six hundred times greater than the justice-level random effect variance. A number of independent variables that are found in the justice-level model to be statistically significant at the 0.05 level are not significant in the case-level model. These are all case-level variables, and include Solicitor General Appellant, Washington Elite Appellant, Law Professor Appellant, and the Difference in Litigating Experience. It appears that these effects were concluded to be significantly different from zero due to specification bias. Also, three of the five variables added to the model – SG Appellee Amicus, SG Appellant Amicus, and Lower Court Conflict – are statistically significant in the expected direction. Evidence for the bloc of added variables is moderate in that the CCVLL is better in the full model, but the BIC is highest in the model that is only extended with a case-level random effect.

The models in table 1 can be used to predict many features of the Court data. I examine the relative performance of the justice and case-level models through their prediction of the size of the voting majority in a case (e.g. 9-0, 8-1, 7-2 etc.). The ropeladder plot in figure

---

8I also considered a model with random effects at both the justice and case levels, but a likelihood ratio test indicates that the justice-level random effect does not improve the model.
3 compares the predicted distribution of the size of the majority to the distribution of the majority sizes over the 443 cases in the actual data. The y-axis gives the majority size, and the horizontal axes give the number of cases out of the 443. The points represented as squares give the predictions from the original model, and the diamonds give the predictions of the Case + model. It can be seen that overall the case-level model provides a much better fit than the original model, and where the improvement is most prevalent is in the tails of the distribution. Where the case-level model accurately predicts provides accurate predictions for all of the majority sizes, the original model does very poorly at predicting majorities of size 5, 6 and 9. Moreover, the modal case-level outcome in the data is a unanimous decision, which occurs in 153 out of the 443 cases. The original model predicts a frequency of unanimous decisions of fifteen. In short, there is a great deal of case-level consensus in voting on the Court, and failure to account for this results in a biased specification which leads to faulty inferences regarding the effect of independent variables as well as predictions regarding case-level outcomes.

[Insert Figure 3 here]

The picture of the Court discovered here is much different than that painted by the dominant attitudinalist perspective on Court behavior, which contends that most of the Court’s voting on the merits is driven by ideology (Segal and Spaeth 2002). An enormous amount of variance exists at the case-level – so much that simply adding the case-level random effect increases the log-likelihood more than all of the covariates combined. The improvement suggested by the joint prediction error analysis (1) demonstrates an empirical failure of the attitudinal model, (2) permits more reliable inferences on the effects of covariates than those published in the original article, and (3) directs attention in the way of past findings that have been inappropriately excluded from the model.
4.2 The Reliability of Democratic Allies

4.2.1 Examination of Original Findings

In the second replication, I examine international defense alliance fulfilment. Gartzke and Gleditsch (2004) test whether democratic allies are more or less likely than non-democratic allies to provide military aid to an ally that is attacked. Their hypothesis is that democratic states, due to the domestic audience costs of military intervention in a conflict involving an ally, are less likely to aid an ally than non-democratic states. To test their hypothesis Gartzke and Gleditsch (2004) study the participation of allies in wars from 1816 to the present. For each war considered, all of the allies of the participants are included in the dataset. The dependent variable is binary; coded 1 if the ally provided military aid and 0 otherwise. A logistic regression model is specified where the main independent variable of interest is an indicator of whether or not the ally has a democratic government (i.e. if the Polity II score is greater than 6). Other control variables include: whether the ally is contiguous to the attacked state, whether the ally is allied to the aggressor, and the COW composite combined capabilities score (CINC) of both the ally and the attacked state. They find that democratic states are less likely than non-democratic states to provide military aid to allies. It is relevant to note that Gartzke and Gleditsch (2004) assume, at least implicitly, that there is no interdependence among those allies considering intervention into the same conflict. This proves to be an inappropriate assumption.

In the JPE analysis, the collective I consider is the group of states considering intervention on the same side of a conflict. There is good reason to expect correlation among states considering intervention in the same side of a conflict. First, there is very little in the way of conflict-specific information in the model, which would induce correlation through unobserved war-level covariates. Examples of potentially important factors that are omitted include whether the assets of third parties are endangered by the conflict (Butler 2003), the
history of interventions in the conflicts of the target state (Gleditsch and Beardsley 2004),
and the number of states involved in the conflict (Kim 1991). Another possibility is that explicit coordination occurs among allies to states in a given conflict. Powerful international institutions such as the North Atlantic Treaty Organization (NATO) and the United Nations (UN) exist in part to coordinate the military intervention activities of member states (Hartley and Sandler 1999; Solana 1999; Lebovic 2004; Lango 2005). Lastly, intervention decisions by individual states are interdependent means to a common end – the result of the conflict. If the U.S. intervenes on behalf of one side of a conflict, Canada may no longer need to intervene to produce a victory for the side receiving help from the U.S.

The parameters of the JPE analysis are set at the same levels as in the Supreme Court example: the number of draws $t = 1,000$, the size of the JPE $k = 2$, and the p-value $\alpha = 0.10$.$^9$ A total of 1,071 JPEs are discovered. All of them are under-predictions, 807 of which are pairs of states making the same intervention decisions. Two interesting patterns emerge. First, since approximately 80% of the under-predictions are states in agreement, it appears the original model underestimates the degree of correlation between states considering assistance to one side of a conflict.

A second pattern in the JPEs regards the types of states that intervene more often than are predicted and those that intervene less often. Examining the list of prediction errors, I noticed a stark difference between two areas of the globe that are less than completely democratic – Latin America and the Middle East. Latin American states intervene in conflicts much less often than predicted and Middle Eastern states intervene much more often than predicted. This is depicted in figure 4, where it is seen that the model in Gartzke and Gled-

$^9$As in the Court example, deviations from these parameter values do not produce different substantive inferences.
itsch (2004) disproportionately underpredicts fulfilment decisions by Middle-Eastern States, and non-fulfilment by Latin American states. In the figure, a circle is placed at the capitol of every member in a prediction error. The darker the color of the circle, the greater the number of intervention prediction-errors in which that state is involved. This regional pattern leads to an additional hypothesis regarding the causes of defense alliance fulfilment.

[Insert Figure 4 here]

As was briefly discussed above in reference to the role of international institutions, states often seek the approval and support of other nations when intervening in a conflict. There is debate regarding the ability of third party consultation to mitigate conflict in the international arena (Fisher and Keashly 1991; Diehl, Druckman and Wall 1998; Wilkenfeld et al. 2003), but the argument and findings presented by Ireland and Gartner (2001) support the hypothesis that the international consultation demands in alliance agreements are enough to discourage states from participating in conflicts. Ireland and Gartner (2001) argue that, in many instances, states will seek the approval of allies before entering into a conflict. In fact, many alliance agreements include pacts that require explicit prior consultation. In their empirical analysis of conflict initiation by European parliamentary governments from 1922 to 1996, Ireland and Gartner (2001) find that a consultation pact reduces the instantaneous hazard of conflict initiation by 85% – an effect that is statistically significant at the 0.05 level. States may be motivated to honor consultation agreements in order to create and maintain a reputation for reliable international commitments. A state’s reputation affects inclusion in future international activities. As Crescenzi (2007, pp. 1) observes, “In international politics, states learn from the behavior of other nations, including the reputations states form through their actions in the international system.” Gibler (2008) finds that states with a reputation for upholding defense alliances are more likely to be included in future alliances and that being allied with strong-reputation allies effectively deters military attacks from
other states. Moreover, a state can damage its reputation for reliable international commitment by ignoring consultation obligations (Kagan 2004; Tucker and Hendrickson 2004; Sandler 2005). Given that international consultation obligations can serve as an obstacle to states’ entry into conflict, in the context of the current application, it would be expected that states with more consultation pacts would be less likely to fulfil defense alliances due to consultation’s constraint on conflict initiation. A comparison of the regional patterns in the consultation alliance network with those in the prediction errors suggests that a state’s consultation obligation is an important omitted variable.

Looking again at figure 4, the size of the point for each state is proportional to the average number of consultation pacts in which it is involved for the years that it appears in the data from Gartzke and Gleditsch (2004). It may seem odd to see a number of small (i.e. poorly connected) states in the heart of Western Europe, but most of these are former German Kingdoms such as Bulgaria. These states appear in the data during conflicts in the 19th century when consultation pacts were not common. The ATOP codebook defines consultation pacts as agreements that, ”obligate members to communicate with one another in the event of crises that have the potential to result in military conflict with the goal of creating a joint response.” (Leeds (2005) p.10). The states with larger points also have lighter points, indicating that better connected states in the consultation network are less likely than predicted by the original model to intervene on behalf of an ally. This pattern is consistent with the hypothesis articulated above – that consultation pacts serve as a hindrance to conflict participation. Given theoretical reasons to expect consultation obligations to matter, if the visual diagnosis presented in figure 4, connectedness in the consultation pact network is an important omitted factor in the explanation of defense alliance fulfilment.
4.2.2 Improved Models of Defense Alliance Fulfilment

I have identified two interesting regularities in the joint prediction errors. First, there seems to be unmodeled positive correlation between the decisions made by the allies of an attacked state. Second, it appears that the consultation obligations of an ally can inhibit the ally from entering into a conflict. Again, we must statistically test whether the patterns discovered in the joint prediction error analysis truly exist in the data generally, and whether accounting for them improves the specification of Gartzke and Gleditsch (2004). To account for correlation among states that are allied to the same state I add a target-conflict random effect to the model, where the target is the state being potentially assisted in the alliance and the conflict is a specific instance of war. To test whether consultation obligations reduce the likelihood of alliance fulfilment, I add a variable to the model (Consultation Degree) which is the number of states with which the ally has consultation pacts in year $t$. If state $A$ must decide whether to intervene into a conflict in year $t$, Consultation Degree is the number of states with which state $A$ has consultation pacts in year $t$. Table 5 presents the results with various specifications that include the improvements identified in the JPE analysis.

[Insert Table 2 here]

The results support the inferences suggested in the JPE analysis. In terms of the first pattern discovered in the JPE analysis, there is a high degree of association between the decisions rendered by states in the same target-conflict group. The addition of a target-conflict random effect improves model fit considerably. Over all three of the covariate specifications, the addition of the target-conflict random effect improves the BIC and CCVLL by 20-30 points. The suspicion that consultation obligation is an important omitted variable is confirmed by the results. Consultation Degree is a statistically significant negative determinant of the probability of alliance fulfilment in all of the different specifications. There is a negative relationship between the number of consultation pacts held by a state and the likelihood
of alliance fulfilment. Accounting for this relationship moves the specification closer to the true data generating process, as evidenced by the CCVLL. Overall, the contributions suggested by the JPE analysis improved the explanation of states’ decisions to fulfil defense alliance obligations.

Another result from the improved specification is that the democracy indicator is no longer statistically significant. Simply adding the random effect to the model eliminates the statistical significance of the democracy indicator. In fact, the best fitting model, according to both the BIC and CCVLL, is the one where a random effect and Consultation Degree is included and the democracy indicator is constrained to have no effect. By improving the model specification, I have shown that the previous inference that democratic states are less likely to fulfil defense alliances is attributable to misspecification bias, and not an actual effect.

5 Conclusion

Political scientists have learned much from the study of repeated collective choice processes, where stable sets of well-known actors repeatedly issue individual decisions that have broad collective implications. A cornerstone of theory regarding repeated, salient interaction is that the actors involved equilibrate to sophisticated and highly interdependent choice strategies. Many contexts of repeated collective choice – roll-call voting in legislatures, decisions on the merits in multi-member courts of appeal, and intervention by nations into conflicts or other emergency events – are characterized by a body of individual actors making micro-level decisions that are aggregated into macro-level outcomes with far-reaching consequences (e.g. law, the interpretation of the law, and the results of conflicts). Since these collective interactions are repeated many times throughout history, an expectation of stable and sophisticated patterns of interdependence among the micro-level decisions is
strongly justified. This poses a challenge to the statistical analysis of micro-level decisions in repeated collective choice data. Namely, if patterns of interdependence are a strong component of the data generating process, common parametric models that are used to analyze this sort of data, such as logistic regression, are misspecified and inferences regarding micro and macro-level factors that drive micro-level choices are suspect due to misspecification bias.

In order to make valid statistical inferences with repeated collective choice data, the model specification must account for the forms of interdependence that characterize repeated collective choice data. It can be incredibly burdensome in terms of both computation and interpretation to specify and estimate a model that is robust to any conceivable form of multivariate dependence among discrete choices. I propose a solution to the problem of interdependence in repeated collective choice data that takes advantage of the wealth of knowledge political scientists hold regarding the observations (e.g. legislators, justices or countries) that is above and beyond that contained in the dataset. I propose that researchers estimate a simple model to start – that which represents the best theoretical specification that can be managed – then examine forms of multivariate deficiency, and iteratively improve the specification to include components hypothesized to account for the model’s failures. Specifically, I introduce the joint prediction error, a collective outcome that is poorly predicted by a model, as a tool for discovering unmodeled forms of interdependence. Theoretical examination of commonalities among and individual characteristics of the membership and actions in these JPEs suggests substantive improvements to the specified model. Additionally, I suggest the use of the cross-validated loglikelihood, an unbiased metric of proximity to the true model, as a tool for judging the validity of the improvements derived from the analysis of JPEs. In two empirical applications I demonstrate the utility of the iterative model improvement procedure I propose. In the JPE analysis of the model from Johnson, Wahlbeck and Spriggs (2006) it is found that there is a very strong unmodeled tendency towards consensus on the
U.S. Supreme Court. This is in contrast with a strictly political view of the Court. Updating the model to account for case-level positive correlation among the votes of the justices strongly improves the fit. Moreover, many inferences made in the original paper are shown to result from specification error. The second replication considers an analysis by Gartzke and Gleditsch (2004) of the likelihood that states fulfil defense alliances. The JPE analysis identifies correlation between states considering intervention on the same side of a conflict. Also, prediction errors are consistent with the importance of states’ consultation obligations, which is omitted from the original specification. Both of the patterns in the JPEs lead to improvements in the empirical model.
References


Collins Jr, Paul M. 2008. *Friends of the Supreme Court: interest groups and judicial decision making.* Oxford University Press, USA.


\[ t = \text{number of desired draws from } f_X(\cdot) \text{ (the PPD)} \]
\[ \hat{\theta} = D \times P \text{ MCMC sample} \]
\[ X = N \times M \text{ sample of observed data} \]
\[ \hat{X} = \text{Sample from the PPD initialized to } \emptyset \]

```
for(i in 1 to t) begin
1. Draw \( \theta^{(i)} \) randomly from the rows of \( \hat{\theta} \)
2. Draw \( X_{new}^{(i)} \) (the same size as \( X \)) from \( l(X_{new}|\theta^{(i)}) \)
3. Store \( X_{new}^{(i)} \) in \( \hat{X} \)
end
```

\( \hat{X} \) now contains \( t \) random draws from \( f_X(\cdot) \)

Figure 1: This figure gives the algorithm used to randomly draw data from a model fit by Bayesian or maximum likelihood methods, using only a sample of parameters from the posterior or asymptotic sampling distribution respectively.
Figure 2: Histograms of the number of cases in which the justice-vote pair occurs over the 1,000 datasets drawn from the model. The solid line is the times that pair occurs in the actual data. The four most frequent under and over predictions are given in the left and right columns respectively. The title gives the last name of the justices and the direction of the vote (R–reverse, A–affirm).
Figure 3: Ropeladder plot demonstrating the fit of the models to the size of the majority in Supreme Court cases. Points give predictions, and bars span 95% confidence intervals.
Figure 4: Under-predictions from the model in Gartzke and Gleditsch (2004). Each point is located at the capitol of the state involved in the JPE. The darker the point, the greater the number of intervention JPEs in which that state is involved. The larger the point, the greater the number of consultation pacts in which the state is involved.
<table>
<thead>
<tr>
<th></th>
<th>Justice Level</th>
<th>Case Level</th>
<th>Case Level +</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>Estimate</td>
</tr>
<tr>
<td>Constant</td>
<td>0.280</td>
<td>0.067</td>
<td>0.556</td>
</tr>
<tr>
<td>Ideological Compatibility with Appellant</td>
<td>0.310+</td>
<td>0.017</td>
<td>0.599+</td>
</tr>
<tr>
<td>Oral Argument Grade</td>
<td>0.205+</td>
<td>0.040</td>
<td>0.391+</td>
</tr>
<tr>
<td>Case Complexity</td>
<td>0.075</td>
<td>0.101</td>
<td>0.169</td>
</tr>
<tr>
<td>Oral Argument Grade × Case Complexity</td>
<td>-0.089</td>
<td>0.091</td>
<td>-0.289</td>
</tr>
<tr>
<td>Ideological Compatibility × Oral Argument Grade</td>
<td>0.020</td>
<td>0.016</td>
<td>0.026</td>
</tr>
<tr>
<td>U.S. Appellant</td>
<td>0.472+</td>
<td>0.117</td>
<td>0.914+</td>
</tr>
<tr>
<td>U.S. Appellee</td>
<td>-0.790+</td>
<td>0.150</td>
<td>-1.633+</td>
</tr>
<tr>
<td>S.G. Appellant</td>
<td>0.325+</td>
<td>0.127</td>
<td>0.544</td>
</tr>
<tr>
<td>S.G. Appellee</td>
<td>-0.208</td>
<td>0.167</td>
<td>-0.321</td>
</tr>
<tr>
<td>Washington Elite Appellant</td>
<td>0.406+</td>
<td>0.136</td>
<td>0.765</td>
</tr>
<tr>
<td>Washington Elite Appellee</td>
<td>0.069</td>
<td>0.145</td>
<td>0.110</td>
</tr>
<tr>
<td>Law Professor Appellant</td>
<td>-0.757+</td>
<td>0.269</td>
<td>-1.283</td>
</tr>
<tr>
<td>Law Professor Appellee</td>
<td>-1.554+</td>
<td>0.323</td>
<td>-3.007+</td>
</tr>
<tr>
<td>Clerk Appellant</td>
<td>-0.246</td>
<td>0.154</td>
<td>-0.571</td>
</tr>
<tr>
<td>Clerk Appellee</td>
<td>-0.165</td>
<td>0.197</td>
<td>-0.145</td>
</tr>
<tr>
<td>Elite Law School Appellant</td>
<td>0.025</td>
<td>0.088</td>
<td>0.090</td>
</tr>
<tr>
<td>Elite Law School Appellee</td>
<td>-0.127</td>
<td>0.089</td>
<td>-0.290</td>
</tr>
<tr>
<td>Difference in Litigating Experience</td>
<td>-0.127+</td>
<td>0.034</td>
<td>-0.234</td>
</tr>
<tr>
<td>Appellee Amicus</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Appellant Amicus</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SG Appellee Amicus</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SG Appellant Amicus</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lower Court Conflict</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Justice-Level Variance</td>
<td>0.010</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Case-Level Variance</td>
<td>-</td>
<td>-</td>
<td>6.88</td>
</tr>
<tr>
<td>CCVLL (No RE, RE)</td>
<td>-2,021</td>
<td>-2,019</td>
<td>-2,064</td>
</tr>
<tr>
<td>BIC</td>
<td>4,153</td>
<td>3,253</td>
<td>3,274</td>
</tr>
<tr>
<td>N</td>
<td>3,331</td>
<td>3,331</td>
<td>3,331</td>
</tr>
<tr>
<td>Clusters</td>
<td>16</td>
<td>443</td>
<td>443</td>
</tr>
</tbody>
</table>

Table 1: U.S. Supreme Court voting on the merits. Hierarchical logistic regression estimates are presented. + statistically significant at the 0.05 level (one-tailed). The CCVLL is the cluster cross-validated log-likelihood.
<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>RE</th>
<th>CD, ND</th>
<th>CD, ND, RE</th>
<th>CD</th>
<th>CD, RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-2.16</td>
<td>-3.94</td>
<td>-1.53</td>
<td>-2.89</td>
<td>-1.43</td>
<td>-2.85</td>
</tr>
<tr>
<td></td>
<td>(0.279)</td>
<td>(0.786)</td>
<td>(0.396)</td>
<td>(0.834)</td>
<td>(0.402)</td>
<td>(0.866)</td>
</tr>
<tr>
<td>Consultation Degree</td>
<td>-</td>
<td>-2.53</td>
<td>-0.0523</td>
<td>-0.0701</td>
<td>-0.0483</td>
<td>-0.0698</td>
</tr>
<tr>
<td></td>
<td>(0.0195)</td>
<td>(0.0438)</td>
<td>(0.201)</td>
<td>(0.0436)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A is Democracy</td>
<td>-1.02</td>
<td>-0.108</td>
<td>-0.737</td>
<td>-0.095</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.565)</td>
<td>(0.927)</td>
<td>(0.571)</td>
<td>(0.914)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A Allied to Other Side</td>
<td>-0.0536</td>
<td>-0.355</td>
<td>-0.0549</td>
<td>-0.221</td>
<td>-0.0154</td>
<td>-0.24</td>
</tr>
<tr>
<td></td>
<td>(0.315)</td>
<td>(0.905)</td>
<td>(0.318)</td>
<td>(0.832)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A and B Contiguous</td>
<td>0.911+</td>
<td>1.09+</td>
<td>0.882+</td>
<td>1.05+</td>
<td>0.776+</td>
<td>1.03+</td>
</tr>
<tr>
<td></td>
<td>(0.31)</td>
<td>(0.507)</td>
<td>(0.314)</td>
<td>(0.832)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CINC A</td>
<td>-4.05+</td>
<td>-7.15</td>
<td>-6.75+</td>
<td>-11+</td>
<td>-6.74+</td>
<td>-11+</td>
</tr>
<tr>
<td></td>
<td>(2.45)</td>
<td>(7.11)</td>
<td>(2.8)</td>
<td>(6.72)</td>
<td>(2.82)</td>
<td>(6.7)</td>
</tr>
<tr>
<td>CINC B</td>
<td>7.43+</td>
<td>11.4+</td>
<td>5.76+</td>
<td>11.3+</td>
<td>6.31+</td>
<td>11.3+</td>
</tr>
<tr>
<td></td>
<td>(2.76)</td>
<td>(5.06)</td>
<td>(2.81)</td>
<td>(4.92)</td>
<td>(2.86)</td>
<td>(4.9)</td>
</tr>
<tr>
<td>Coalition-Level Variance</td>
<td>-</td>
<td>7.15</td>
<td>-</td>
<td>5.69</td>
<td></td>
<td>5.60</td>
</tr>
<tr>
<td>CCVLL</td>
<td>-175.5</td>
<td>-141.7</td>
<td>-173.8</td>
<td>-139.1</td>
<td>-174.4</td>
<td>-140.6</td>
</tr>
<tr>
<td>BIC</td>
<td>351.9</td>
<td>306.4</td>
<td>347.3</td>
<td>303.1</td>
<td>345.4</td>
<td>309.2</td>
</tr>
</tbody>
</table>

Table 2: Results presented are logistic regression coefficients with standard errors in parentheses. + statistically significant at the 0.05 level (one-tailed). Model abbreviations are as follows; RE = Random Effect, CD = Consultation Degree, ND = No Democracy. A total of 451 observations with 91 target-conflict groups are used in each model. The CCVLL is the cluster cross-validated log-likelihood.