Public Choice (2012) 151:719–737
DOI 10.1007/s11127-011-9767-1

Lessons in disguise: multivariate predictive mistakes in collective choice models

Bruce A. Desmarais

Received: 13 November 2010 / Accepted: 11 January 2011 / Published online: 21 January 2011
© Springer Science+Business Media, LLC 2011

Abstract Many political processes can be characterized as repeated collective decisions, in which individual choices rendered by the same actors are combined to produce salient outcomes (e.g., voting in legislatures). The inherent interdependence among the choices within repeated collective decisions is often missed in statistical analysis. I introduce the joint prediction error (JPE)—a grouping of individual decisions that is poorly predicted by a model. JPEs capture the intersecting information missed by conventional diagnostics. I demonstrate the use of JPEs on data from two published articles—one on U.S. Supreme Court voting and another on international defense alliances.

Keywords Collective choice · Prediction · Discrete choice · Supreme Court · Alliance · Interdependence · Correlation · Diagnostic

> "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

Many political processes can be characterized as repeated collective decisions. Repeated collective decisions are recurrent events in which a relatively stable group of actors issue individual choices on the same question; choices which are combined into a single decision/outcome with respect to that question. Examples of this in domestic politics include roll-call voting in legislatures and decisions issued by multi-member courts of appeal. In the international arena, coordinated intervention into civil wars, the provision of relief for natural disasters, and the issuance of trade sanctions are repeated collective decisions taken by

B.A. Desmarais (✉)
Department of Political Science, University of Massachusetts Amherst,
Thompson Hall, 200 Hicks Way, Amherst, MA 01003, USA
e-mail: desmarais@polsci.umass.edu

states. Due to the importance of the collective outcomes that result from individual decisions (e.g., laws created 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 among the decision-makers are likely to emerge in repeated collective choice data.[1] However, in empirical studies of repeated collective choice, the members of the stable group are often pooled into a sample for regression modeling, where relationships between the members are ignored to the extent that they are not captured by independent variables. If such relationships do exist, 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 about and modeling these dependencies. Similar in structure to the approach advocated by Achen (2005), rather than estimate an overly complicated model at the onset, 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 those updates.

Since explicit combinations of the micro-level political choices under study imply the collective outcomes, in order for the micro-level model to be correctly specified, it must account for higher-level tendencies in the subject group. For instance, Hix et al. (2005) find that there are varying levels of political party cohesion in the European Parliament. If the findings of Hix et al. (2005) are valid, any micro-level model 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, and can be challenging to interpret.[2] Rather than attempt to extend a micro-level model to accommodate every form of interdependence that has either found support in the literature or can be reasonably conceived, I develop a method to identify configurations of individual decisions that contradict the dependence structure of a baseline model. The theoretical analysis of these joint outcomes suggests the key multivariate extensions necessary to make valid inference on the micro-level processes. Analogous to diagnosing temporal dependence in time series data, the method I propose is a diagnostic tailored to dependence features likely to characterize repeated collective choice data.

Generally speaking, residual analysis involves the comparison of observed data with predicted values from a statistical model, with the goal of identifying major inconsistencies between the model and the data. Commonly applied forms of residual analysis include diagnostics for serial correlation and heteroscedasticity. Just as the Durbin-Watson test (Durbin and Watson 1950) was designed to diagnose a common problem in time series data, I argue that dependence among decision-makers characterizes collective choice data, and introduce a flexible diagnostic tool—a *joint prediction error* (JPE). A JPE is a collective outcome that is observed to occur with a much different frequency than predicted by a given statistical model.

[1] 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).

[2] See, e.g., Alvarez and Nagler (1998) for an example where preferences for electoral candidates are posited to be correlated, Franzese et al. (2007) for a discussion of the estimation challenges in accounting for spatial dependence in time-series cross-section data, and Ward et al. (2007) who find that latent reciprocal and transitive tendencies characterize international dyadic data.

By replicating and extending two recently published studies, I demonstrate how improvements in models of repeated collective discrete 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 et al. (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 of international defense alliance activation, 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.

2 Joint prediction errors

Rubin (1984) noted that frequency calculations performed on real data should not differ from model predictions in relevant ways. Since, in a repeated collective choice setting, there are many opportunities for the dependence among choices to emerge, the relationships among the decision-makers are of great relevance to the process of correctly specifying a statistical model. Thus, I define the joint prediction error as a combination of choices that occurs with a much different frequency than that predicted by a model. To avoid misspecification bias, whenever possible, a model should be updated to eliminate JPEs.

The specific metric used to determine whether a joint outcome constitutes a JPE is the *posterior predictive p-value* introduced by Meng (1994) for general use in a Bayesian context. This p-value can be used to assess the oddity of the frequency of a joint outcome given predictions from a model. For instance, it would allow one to test whether the frequency of unanimous decisions on the U.S. Supreme Court is statistically significantly different from the frequency of unanimous decisions predicted by a statistical model. The posterior predictive p-value is defined with regard to some function of the data $T(X)$. The objective is to determine whether the observed value of $T(X)$ is an outlier with respect to the predicted distribution of $T(X)$. This predicted distribution of $T(X)$ is derived by drawing many simulated versions of the data from the posterior predictive distribution of the data implied by the fitted model. Let $p(\theta|X)$ be the posterior distribution of the parameters, and $l(X|\theta)$ the likelihood of the data given the parameters, then the posterior predictive distribution of the new (simulated) data is

$$f(X_{new}) = \int_{\Theta} l(X_{new}|\theta)p(\theta|X)d\theta. \tag{1}$$

King et al. (2000) show how to simulate new data from a model fit within the classical framework (i.e., maximum likelihood) that approximates data simulated from the Bayesian posterior predictive distribution. The oddity of the observed value of $T(X)$ is given by the lesser of the proportion of simulated values of $T(X)$ smaller than the observed value and the proportion of the simulated values of $T(X)$ greater than the observed value. This proportion gives the (one-sided) posterior predictive p-value. An observed value of $T(X)$ that is either greater than or less than a large proportion of the simulated values indicates that the model

specification does not adequately represent the feature of the data generating process that accounts for $T(X)$.

Given a dataset in which N collective choices occur, the first step in identifying JPEs is to define $T(X)$ as the number of times a particular joint event occurs within those collective choices. A general way to approach this is to define the $T(X)$ to be all decision-maker choice combinations of size k (i.e., if $k = 2$, for every pair of decision-makers, the number of times both said Yea, the number of times both said Nay, as well as the number of times the first (second) said Yea and the second (first) said Nay. Two other parameters must be defined in order to identify JPEs—the number of draws from the posterior predictive distribution (t), and the p-value at which a joint outcome will be judged to be predicted with error (α). It is important to set t high enough to mitigate simulation error, which should be diagnosed by re-running the JPE procedure to assure that the same JPEs are identified. The right value for α will be that which permits significant insight into model improvements—not so high that no JPEs are missed, and not so low that unsuccessful model improvements are tried. Note that the parameters can be tuned liberally, since any insights gathered from the JPEs will be tested with rigorous scrutiny, which I address in Sect. 5.

3 Learning from joint prediction errors

There are two reasons in particular to expect theoretical innovations to arise through the inspection of JPEs in the study of repeated collective choice. First, 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. Second, there is likely to be an overwhelming amount of previous theoretical and empirical research that precedes any new study of historical political data—meaning there are likely to be numerous omissions in any new initial model. Both of these features present unique opportunities for improvement with joint prediction error analysis.

Scholars of repeated collective choice typically have rich historical knowledge of the observations under study. In their analysis of the representational efficacy of majority-minority congressional districts, Cameron et al. (1996, 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*. 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 for collecting and including in the model a measure of a state's development. This auxiliary information optimizes potential benefits from simply examining those combinations of cases 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 on politics, 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. Indeed, such a model would likely lead to a convoluted interpretation, and be counter to the objective of data reduction (Achen 2005). At the same time, previous findings cannot be ignored simply for the sake of time or parsimony. Examining joint prediction errors constitutes a 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 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, and, thus, whether seniority should be added to the model.

4 The mechanisms underlying joint prediction errors

JPE analysis is a model specification diagnostic tool designed with the structure and potential challenges of collective choice data in mind. Specifically, JPEs provide evidence that the model does not adequately represent sources of association between a group's individual decision-makers. Just as tests for autocorrelation and heteroscedasticity can identify the problem, but not necessarily the source (e.g., autocorrelation can be caused by actual memory in the outcome variable or by the omission of an autocorrelated covariate), there are many processes that may lead to the discovery of JPEs. In this section, I discuss some common patterns that might arise and suggest model improvements that would account for those features in the data. It is assumed that the structure of the data does not permit the identification of the sequence of decisions that comprise a particular collective choice. The modeling suggestions conform to this data structure. However, the discovery of JPEs may serve as motivation for collecting decision-timing data in order to unpack the dynamics of association. An excellent example of the added leverage provided by the timing of decisions, in the context voting on the U.S. Supreme Court, is given by Johnson et al. (2005).

As is the case with the applications I present below, it could be that the model underestimates the degree of consensus among the group's members. In this instance, a large number of JPEs in which the decision-makers are in agreement will be discovered. Two mechanisms that could give rise to this are that (1) there is an omitted choice-level covariate in the model and (2) there is consensus-building in the collective (i.e., those in the majority coalition succeed at persuading those in the minority to join the majority). Either of these processes will result in a greater degree of positive association among those who participate in making collective decisions than that accommodated by a conventional discrete choice regression model. The addition of a collective-choice-level random effect to the model will update the structure of the model to accommodate exchangeable positive correlation among the decision-makers (Gelman and Hill 2007). This will improve the estimates of the regression coefficients in the model and relax the assumption that the individual choices are independent, but there are limitations to just adding a random effect. Notably, the incorporation of a random effect assumes that the omitted factor is unrelated to any of the other covariates in the model. Thus, whenever possible, choice-level covariates that have found support in the literature should be incorporated into the updated specification.

The JPEs could reveal a variety of positive and negative associations. This could arise from either (1) the omission of a decision-maker (i.e., individual) level covariate or (2) the development of a variety of influential and conflictive relationships within the group. For

instance, suppose that legislators' preferences are partly determined by the density of the population in the district (i.e., representatives of farmers tend to vote differently than those of urbanites). If a measure of density is not included in the model, legislators representing very similarly dense districts will be positively correlated and those with very different densities will be negatively correlated. Models designed to accommodate unconstrained forms of association, such as the multivariate probit regression (Chib and Greenberg 1998), could be employed to update the model to represent the dependence among decision-makers. However, since the omitted factors might be correlated with those already included in the model, it would be ideal to consider additional covariates that have either been proposed in past work or are justified theoretically.

These are just a couple of patterns in the JPEs that might be discovered. The innovative use of random effects or correlated choice models will prove useful in relaxing the model-based assumption that individual decisions are independent. This offers the benefit of making more valid inferences on the covariates included in the model, but does not necessarily enhance the substantive understanding of the process under study. Close examination of the JPEs will hopefully lead to theoretical innovation regarding the inadequacies of common conceptions of the process. This may require the collection of additional covariates. In the application to international defense alliance fulfillment below, examination of the JPEs leads to a novel conception of the role of international agreements in the fulfillment of defense treaties. Also, it may require the collection of information on the sequence of individual decisions. Lastly, I note that improving the model based on patterns in the JPEs is only an intermediate step. These proposed improvements must be rigorously evaluated in order to avoid overfitting the data.

5 Avoiding overfitting

The process of improving model specification via JPE analysis is clearly inductive in that it is a method designed to resolve inconsistencies between the distribution implied by the model and the empirical distribution of the data. Perhaps the most notable criticism of inductive methods is that they have the potential to overfit the data (Jensen and Cohen 2000). When a model is overfit, ideosyncratic sampling error is attributed to structure in the data generating process (e.g., a spurious correlation leads to the inclusion of a covariate in the model). As such, it is important that JPE-suggested improvements be applied and evaluated in a way that effectively avoids overfitting. The key to avoiding overfitting is to evaluate the performance of a model using data that were *not* used to estimate that model's parameters. That way, structure attributed to noise in the estimation (training) data will not contribute to the model's performance index, which is derived from the evaluation (validation) data. In considering precisely the problem of inductively selecting one statistical model among multiple models under consideration, Jensen and Cohen (2000) show that the method of cross-validation can be used to avoid overfitting in the context of inductive model selection.

The first step in cross-validation is to partition the data into m disjoint samples, which is typically done at random. The parameters are then estimated m times, each time leaving one of the samples out of the estimation process. Given a measure of model fit, the cross-validated performance is then computed by combining (typically summing) the fit measure computed on each sample using the parameters estimated without that sample. In the context of a linear regression model, a variant of the cross-validated squared error (Hjorth 1993) would be defined by computing the familiar R^2 for each of the m samples, using the regression coefficients estimated on the data constituted by the other $m - 1$ samples.

Two critical choices in cross-validation are (1) the method used to split the sample and (2) the fit measure that is cross-validated. Van der Laan et al. (2004) provide strongly advisable suggestions regarding both of these choices. The fit measure is the cross-validated log-likelihood (CVLL). Let $f(\boldsymbol{x}, \boldsymbol{\theta})$ be the probability density or mass function that defines the model, then

$$\text{CVLL} = \sum_{i=1}^{m} \ln\left[f(\boldsymbol{x}_i, \boldsymbol{\theta}^{-i})\right], \tag{2}$$

where $\boldsymbol{x}_i$ is the ith sample in the m-wise partition and $\boldsymbol{\theta}^{-i}$ is the parameter estimate omitting $\boldsymbol{x}_i$ from the estimation. Van der Laan et al. (2004) show that, for a general class of data-splitting methods, the CVLL provides a finite-sample unbiased estimate of the Kullback-Leibler divergence (KLD) of $f(\cdot)$, and that the CVLL converges asymptotically to a sample-independent estimator of the KLD that depends on knowledge of the model that generated the data. This means that the CVLL is a measure of model fit that behaves as if it were computed on data that were (1) drawn from the same distribution as the data used to estimate the model and (2) were not actually used to estimate the model. The Kullback-Leibler Divergence between the estimated model $f(\cdot)$ and the model that generated the data $g(\cdot)$, is a common entropy-based measure of distance between two probability density or mass functions. Consistent estimation of the KLD is, for instance, the motivation for the derivation of Akaike's (1974) information criterion. Thus, the CVLL will not favor overfitting, because fitting sampling ideosyncracies will move the estimated model further from the one that generated the data (van der Laan et al. 2004).

In order for the CVLL to retain the property that it behaves as if it were estimated on an independent sample from the distribution that generated the data, van der Laan et al. (2004) note that the data-splitting scheme must be designed such that the size of each of the m validation samples approaches infinity as the sample size approaches infinity. Any data-splitting scheme where m is fixed (i.e., each validation sample is of size N/m) meets this criterion. Importantly, this excludes leave-one-out cross-validation, which provides an unbiased estimate of the KLD (Smyth 2000), but does not converge to an estimator with knowledge of the model that generated the data. A common form of cross-validation that meets this criteria is 10-fold cross-validation, where the data are randomly split into 10 validation samples (Jensen and Cohen 2000), and this is the method I use to evaluate model improvements suggested by JPE analysis.[3]

Cross-validation of the likelihood function has proven successful in a number of problems where in-sample fit is always improved by adding additional components to the model, yet the analyst either (1) knows that there should be a limited number of features of the model or (2) can only make use of a reasonably parsimonious model. Smyth (2000) shows that the CVLL can be successfully used to select the number of clusters in probabilistic clustering models. Smyth (2000) proves that, on average, selection based on the CVLL will lead the analyst to choose the number of clusters that actually exist in the data, despite the fact that the in-sample fit of a probabilistic clustering model is maximized when the number of clusters is equal to the number of observations in the sample. In a related problem, Knafl and Gray (2007) show that the CVLL can be used to select the number of factors in factor-analytic models—models in which in-sample fit always increases by increasing the number of factors. Additionally, Horne and Garton (2006) show that likelihood cross-validation is also effective for selecting the bandwidth in kernel density estimation, where the in-sample

[3]The model evaluations presented below are robust to varying m from five to 20.

fit is maximized as the bandwidth limits to zero, stacking all of the density or mass on the values in the sample. Given the CVLL's general theoretical properties and its ability to avoid overfitting in a variety of high-dimensional selection problems, it is an ideal metric for preventing overfitting in JPE analysis.

6 Replications with JPE-suggested extensions

6.1 The U.S. Supreme Court and oral argument quality

Johnson et al. (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 to 1994 terms. Johnson et al. (2006) specify a logistic regression model of votes (pooled over justices, cases and terms) where the dependent variable is coded one if the justice votes to reverse the lower court decision and zero for affirm. The votes of Justice Blackmun are excluded due to concerns about endogeneity.[4]

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 two of Table 3 in Johnson et al. (2006). I used $t = 5{,}000$ 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.[5] Figure 1 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 5,000 datasets drawn from the

[4] A number of other control variables are included. The *Difference in Litigating Experience* is the difference in the number of times the appellee's and appellant's attorneys previously argued before the U.S. Supreme Court. *S.G. Appellee* and *S.G. Appellant* are indicator variables for whether the Solicitor General argued for the appellee or the appellant in the case, respectively. *U.S. Appellant* and *U.S. Appellee* are indicator variables for whether a federal government attorney argued the side of the appellant and appellee, respectively. *Elite Law School* is an indicator for whether the respective attorney attended Harvard, Yale, Columbia, Stanford, Chicago, Berkeley, Michigan, or Northwestern Law. *Washington Elite* is an indicator for whether the respective attorney's office is in Washington DC, excluding federal government attorneys. *Law Professor* is an indicator for whether the respective attorney is a law professor. *Clerk* is an indicator for whether the respective attorney was ever a clerk for the U.S. Supreme Court. *Ideological Compatibility* is constructed as the Martin-Quinn score of the justice if the lower court made a conservative ruling and as the negative Martin-Quinn score of the justice if the lower court made a liberal ruling (assuming the petitioner seeks a ruling in the opposite direction of the lower court's decision). *Case Complexity* is constructed using a factor analysis of the number of legal provisions in a case and the number of issues involved in the case. See the original article for the justifications for including these variables.

[5] 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 $t = 5{,}000$ is sufficiently large to avoid simulation error. Also, the substantive inferences I draw from the JPE analysis do not change for α 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.

original model in Johnson et al. (2006). The number of cases in which the pair occurs in the actual dataset is located at the solid vertical line in each panel.[6]

Examining Fig. 1 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 occurs on the Court. Previous studies have found that the Court tends towards consensus decision-making (Haynie 1992; Epstein et al. 2001). It could also be that omitted legal factors are producing correlation (Spriggs et al. 2001; Collins 2004; Johnson et al. 2006). 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. Consensus prediction errors do not constitute a statistical test for the presence of unobserved association in justices' votes. In order to perform a principled test of the intuitions gathered from the JPE analysis, and asses the impact of these patterns on other inferences from the model, the model from Johnson et al. (2006) must be improved to both test and account for positive case-level correlation among the justices.

6.1.1 Case-level determinants of Supreme Court votes

I extend the model in Johnson et al. (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) shows that the Court responds to *Amicus Curiae* briefs. Specifically, he shows that the probability that a particular side wins a case is directly proportional to the number of briefs filed on its behalf. Moreover, briefs filed by the U.S. Solicitor General have a larger effect on the Court's decisions than do those filed by others. The variables *Appellee Amicus*, *Appellant Amicus*, *SG Appellee Amicus* and *SG 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.[7]

[6] R package `Arules` (Hashler et al. 2009) was used to perform the frequent item-set mining. I do not replicate the model in column one of Table 3 in Johnson et al. (2006) because an LR test strongly rejects the hypothesis that the restrictions in the reduced model are valid.

[7] The data for the added controls come from replication data for the analyses in Collins (2008) made available on Paul Collins' website at http://www.psci.unt.edu/~pmcollins/data.htm.

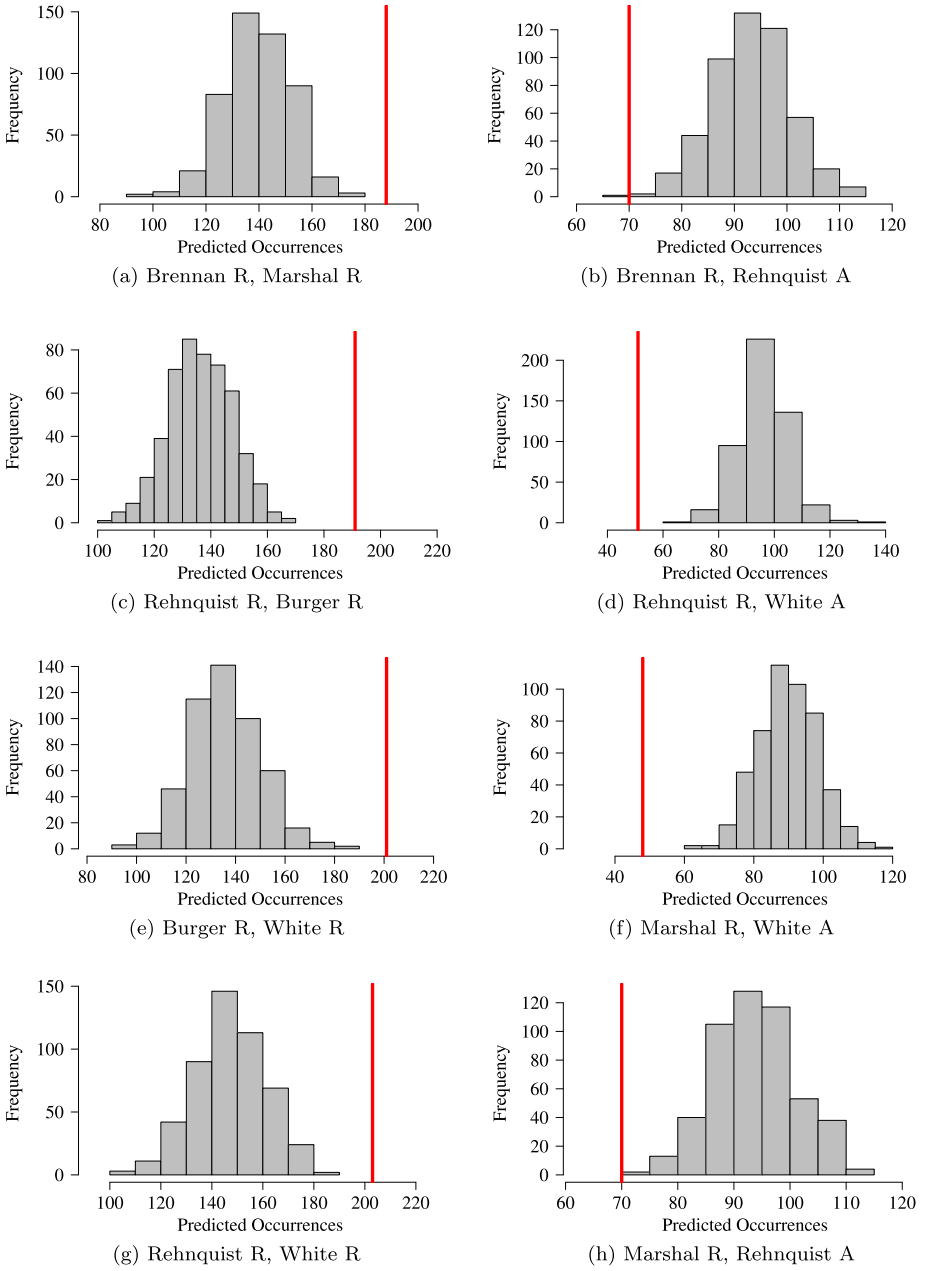

Fig. 1 Histograms of the number of cases in which the justice-vote pair is predicted to occur; with the solid line located at the actual number of occurrences. The title gives the last name of the justices and the direction of the vote (R—reverse, A—affirm)

It would be overly optimistic to assume that all of the case-level association discovered in the JPE analysis would be explained by the covariates I add to the model. I update the model explicitly to 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 for every observation in a cluster. The shared disturbance, assumed to be normally distributed with zero mean, 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 et al. 2007). Thus, the second update to the model presented in Johnson et al. (2006) is to add a case-level random effect.

The results of the hierarchical logistic regression models are presented in Table 1.[8] The model closest to the baseline specification that appeared in Johnson et al. (2006) is the Justice-Level specification. Johnson et al. (2006) use cluster-robust standard errors 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 heterogeneity (Greene 2008, 517), 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.[9] The CVLL is computed by 10-fold cross-validation, randomly splitting the data into 10 samples of approximately 44 cases each.

The pattern discovered in the joint prediction error analysis led to a specification that greatly improves out-of-sample 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 CVLL and BIC by almost 25%. Also there is much more unobserved heterogeneity and/or correlation at the case-level than at 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 CVLL is better in the full model, but the Bayesian information criterion (BIC) is lowest in the model that is only extended with a case-level random effect. Another important finding is that the coefficient on *Oral Argument Grade*—the variable used to test the primary theoretical proposition in the original article—nearly doubles in size in the updated models from 0.205 to 0.391 and 0.40.

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, 6-3, 5-4). The ropeladder plot in Fig. 2 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. It can be seen that the case-level model predicts majority coalition sizes much more accurately than the original model, and where the improvement is most prevalent is in the tails of the distribution. Where the case-level model provides accurate predictions for all of the majority sizes, the original model does very poorly at predicting majorities of size five, six and nine. Moreover, the

[8] R package *lme4* (Bates and Sarkar 2006) was used to estimate the models in Table 1.

[9] I 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.

Table 1 U.S. Supreme Court justices' votes on the merits

	Justice Level		Case Level		Case Level +	
	Estimate	SE	Estimate	SE	Estimate	SE
Constant	0.280	0.067	0.556	0.214	0.78	0.24
Ideological Compatibility	0.310^{+}	0.017	0.599^{+}	0.027	0.599^{+}	0.0265
Oral Argument Grade	0.205^{+}	0.040	0.391^{+}	0.141	0.400^{+}	0.138
Case Complexity	0.075	0.101	0.169	0.366	0.137	0.359
OA Grade × Case Complexity	−0.089	0.091	−0.289	0.306	−0.252	0.301
Ideo. Compatibility × OA Grade	0.020	0.016	0.026	0.025	0.026	0.025
U.S. Appellant	0.472^{+}	0.117	0.914^{+}	0.416	1.17^{+}	0.447
U.S. Appellee	-0.790^{+}	0.150	-1.633^{+}	0.544	-1.83^{+}	0.553
S.G. Appellant	0.325^{+}	0.127	0.544	0.447	0.096	0.485
S.G. Appellee	−0.208	0.167	−0.321	0.599	0.164	0.607
Washington Elite Appellant	0.406^{+}	0.136	0.765	0.483	0.499	0.478
Washington Elite Appellee	0.069	0.145	0.110	0.516	0.312	0.513
Law Professor Appellant	-0.757^{+}	0.269	−1.283	0.957	−1.53	0.940
Law Professor Appellee	-1.554^{+}	0.323	-3.007^{+}	1.135	-2.75^{+}	1.11
Clerk Appellant	−0.246	0.154	−0.571	0.541	−0.490	0.531
Clerk Appellee	−0.165	0.197	−0.145	0.690	−0.248	0.684
Elite Law School Appellant	0.025	0.088	0.090	0.316	0.014	0.310
Elite Law School Appellee	−0.127	0.089	−0.290	0.321	−0.342	0.315
Difference in Litigating Experience	-0.127^{+}	0.034	−0.234	0.122	-0.274^{+}	0.120
Appellee Amicus	–	–	–	–	−0.039	0.073
Appellant Amicus	–	–	–	–	−0.027	0.085
SG Appellee Amicus	–	–	–	–	-1.44^{+}	0.559
SG Appellant Amicus	–	–	–	–	1.05^{+}	0.522
Lower Court Conflict	–	–	–	–	-0.946^{+}	0.413
Justice-Level Variance	0.010	–	–	–	–	–
Case-Level Variance	–	–	6.88	–	6.52	–
CVLL	−2,071.74		−1,559.27		−1,557.71	
BIC	4,153		3,253		3,274	
N	3,331		3,331		3,331	
Clusters	16		443		443	

Hierarchical logistic regression estimates are presented. $^{+}$ Statistically significant at the 0.05 level (one-tailed). The CVLL is the 10-fold cross-validated log-likelihood

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 15. 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.

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 the independent, individual ideological predispositions of the justices (Segal and Spaeth 2002). An enormous amount of variance exists at the case-

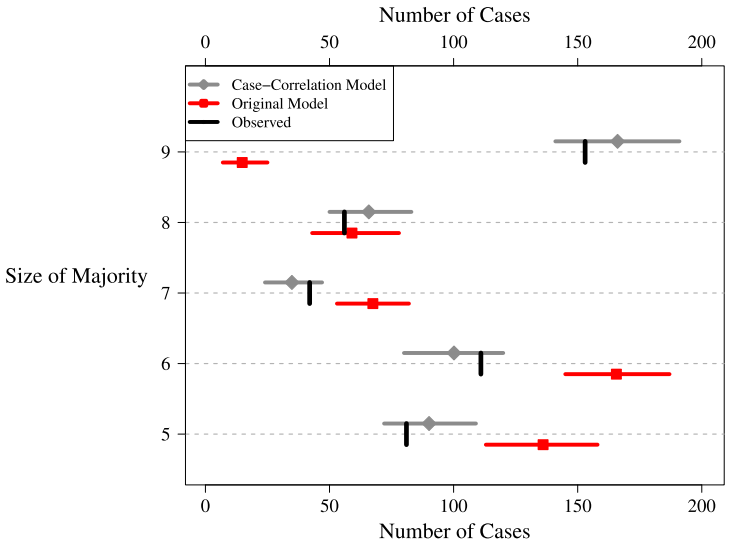

Fig. 2 Ropeladder plot demonstrating the fit of the models to the size of the majority in Supreme Court cases. Points give predictions, and bars span 95% confidence intervals

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 that case-level factors, downplayed in the attitudinal model, are indeed important, (2) permits more reliable inferences on the effects of covariates than those published in the original article, and (3) directs attention towards past findings in the literature (i.e., control variables) that have been inappropriately excluded from the model.

6.2 The reliability of democratic allies

6.2.1 Examination of original findings

In the second replication, I examine international defense alliance fulfillment. 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 one if the ally provided military aid and zero otherwise. They specified a logistic regression model 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 six). Other control variables include: whether the ally is contiguous to the attacked state, whether the ally is allied to the aggressor, and the Correlates of War (COW) composite index of national capabilities (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.

In the JPE analysis, the collective I consider is the group of states considering intervention on the same side of a conflict. This choice is far from arbitrary. First, there

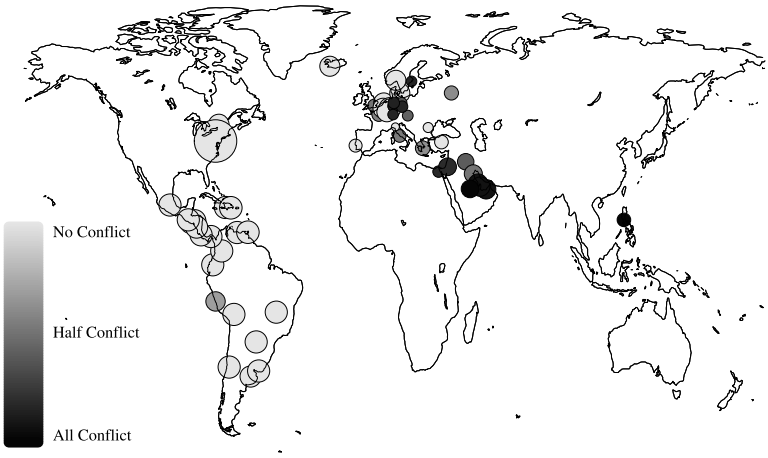

Fig. 3 Under-predictions from the model in Gartzke and Gleditsch (2004). Each point is located at the capitol of the state involved in the JPE. The darker the point, the greater the number of intervention JPEs in which that state is involved. The larger the point, the greater the number of consultation pacts in which the state is involved

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). Lastly, intervention decisions by individual states are interdependent means to a common end—the result of the conflict. If the United States intervenes on behalf of one side of a conflict, Canada may no longer need to intervene for the side receiving U.S. help.

The parameters of the JPE analysis are set at the same levels as in the Supreme Court example: the number of draws $t = 5{,}000$, the size of the JPE $k = 2$, and the p-value $\alpha = 0.10$.[10] 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, there is a clear 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 Fig. 3, where it is seen that the model in Gartzke and Gleditsch (2004)

[10] As in the Court example, deviations from these parameter values do not produce different substantive inferences.

disproportionately underpredicts fulfillment decisions by Middle-Eastern States, and non-fulfillment 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. Middle-Eastern states constitute the largest collection of conflict-prone dark states on the map, and Latin America is a collection of conflict-averse lighter states. This regional pattern leads to an additional hypothesis regarding the causes of defense alliance fulfillment.

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 et al. 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, 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 (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 fulfill 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 Fig. 3, 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).[11] The Alliance Treaty Obligations and Provisions (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, 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, the apparent association between conflict participation and

[11] 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.

Table 2 The reliability of allies and international conflict

	Original	RE	CD, ND	CD, ND, RE	CD	CD, RE
Constant	−2.16	−3.94	−1.53	−2.89	−1.43	−2.85
	(0.279)	(0.786)	(0.396)	(0.834)	(0.402)	(0.866)
Consultation Degree	–	–	−0.0523^{+}	−0.0701^{+}	−0.0483^{+}	−0.0698^{+}
			(0.0195)	(0.0438)	(0.201)	(0.0436)
A is Democracy	−1.02^{+}	−0.108			−0.737	−0.095
	(0.565)	(0.927)	–	–	(0.571)	(0.914)
A Allied to Other Side	−0.0536	−0.355	0.0549	−0.221	−0.0154	−0.24
	(0.315)	(0.905)	(0.314)	(0.828)	(0.318)	(0.832)
A and B Contiguous	0.911^{+}	1.09^{+}	0.882^{+}	1.05^{+}	0.776^{+}	1.03^{+}
	(0.31)	(0.507)	(0.307)	(0.48)	(0.314)	(0.49)
CINC A	−4.05^{+}	−7.15	−6.75^{+}	−11^{+}	−6.74^{+}	−11^{+}
	(2.45)	(7.11)	(2.8)	(6.72)	(2.82)	(6.7)
CINC B	7.43^{+}	11.4^{+}	5.76^{+}	11.3^{+}	6.31^{+}	11.3^{+}
	(2.76)	(5.06)	(2.81)	(4.92)	(2.86)	(4.9)
Coalition-Level Variance	–	7.15	–	5.69	–	5.60
CVLL	−176.3	−141.75	−174.5	−139.98	−173.83	−140.79
BIC	351.9	306.4	347.3	303.1	345.4	309.2

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 CVLL is the 10-fold cross-validated log-likelihood

consultation obligations in the JPEs suggests that the model of alliance fulfillment should account for the connectedness of a state—the expectation being that better-connected states will be less apt to fulfill alliance-based conflict obligations.

6.2.2 Improved models of defense alliance fulfillment

I have identified two interesting regularities in the JPEs. First, it appears that the consultation obligations of an ally can inhibit the ally from entering into a conflict. Second, there seems to be unmodeled positive correlation between the decisions made by the allies of an attacked state. Again, we must statistically test whether the patterns discovered in the JPE analysis truly exist in the data generally, and whether accounting for them improves the specification of Gartzke and Gleditsch (2004). To test whether consultation obligations reduce the likelihood of alliance fulfillment, 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. To account for correlation among states that are allied with 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. Table 2 presents the results with various specifications that include the improvements identified in the JPE analysis.

The results support the inferences suggested in the JPE analysis. In this analysis, the CVLL is constructed by randomly splitting the sample into 10 sub-samples of approximately

nine target-conflict groups each. 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 CVLL by 20–30 points. The suspicion that consultation obligation is an important omitted variable is also confirmed by the results. *Consultation Degree* is a statistically significant negative determinant of the probability of alliance fulfillment in all of the different specifications. Accounting for this relationship moves the specification closer to the true data generating process, as evidenced by the CVLL. Overall, the contributions suggested by the JPE analysis improved the explanation of states' decisions to fulfill 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 CVLL, 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 fulfill defense alliances is attributable to misspecification bias, and not an actual effect.

7 Conclusion

Interdependence among decision-makers is a common feature in repeated collective choice data. Two general phenomena lead to this property. First, choice level, or individual level covariates that are omitted from a statistical model will lead to residual association between individual decisions within the group. Second, repeated interaction can lead to influence and coalition-building among group members. I present the joint prediction error as a tool for diagnosing features of the data that contradict the dependence structure represented by a given statistical model. Analysis of the JPEs leads to updated models that more accurately characterize the dependence among decision-makers in repeated collective choice settings, which, in turn, produces more valid inferences on the process under study. Through rigorous out-of-sample tests, analysts can assure that JPE-suggested extensions move the final model closer to the underlying data generating process rather than overfit the data. Applications to voting on the U.S. Supreme Court and the activation of international defense alliances demonstrate how JPE analysis can suggest substantial improvements to previously published specifications.

Acknowledgements This work was supported in part by the American Politics Research Group, Department of Political Science, University of North Carolina at Chapel Hill and by the College of Social and Behavioral Sciences at the University of Massachusetts Amherst. Without implicating, I would like to thank Tom Carsey, Jim Stimson, Skyler Cranmer, Isaac Unah, Kevin McGuire, Jeff Harden and Justin Kirkland for helpful feedback on this research. Also, the anonymous reviewers' and editors' comments and suggestions were very helpful.

References

Achen, C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, *22*(4), 327.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, *19*(6), 716–723.
Alvarez, R. M., & Nagler, J. (1998). When politics and models collide: estimating models of multiparty elections. *American Journal of Political Science*, *42*(1), 55–96.
Bates, D., & Sarkar, D. (2006). *lme4: linear mixed-effects models using S4 classes*. R package version 0.9975-10.
Butler, M. J. (2003). U.S. military intervention in crisis, 1945–1994: an empirical inquiry of just war Theory. *The Journal of Conflict Resolution*, *47*(2), 226–248.
Caffo, B., An, M.-W., & Rohde, C. (2007). Flexible random intercept models for binary outcomes using mixtures of normals. *Computational Statistics and Data Analysis*, *51*(11), 5220–5235.
Cameron, C., Epstein, D., & O'Halloran, S. (1996). Do majority-minority districts maximize substantive black representation in congress? *The American Political Science Review*, *90*(4), 794–812.
Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361.
Collins, P. M. J. (2004). Friends of the court: examining the influence of *amicus curiae* participation in U.S. Supreme Court litigation. *Law & Society Review*, *38*(4), 807–832.
Collins, P. M. Jr. (2008). *Friends of the Supreme Court: interest groups and judicial decision making*. London: Oxford University Press.
Crescenzi, M. J. C. (2007). Reputation and interstate conflict. *American Journal of Political Science*, *51*(2), 382–396.
Diehl, P. F., Druckman, D., & Wall, J. (1998). International peacekeeping and conflict resolution: a taxonomic analysis with implications. *The Journal of Conflict Resolution*, *42*(1), 33–55.
Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression. I. *Biometrika*, *37*(3/4), 409–428.
Epstein, L., Segal, J., & Spaeth, H. (2001). The norm of consensus on the U.S. Supreme Court. *American Journal of Political Science*, *45*(2), 362–377.
Fisher, R. J., & Keashly, L. (1991). The potential complementarity of mediation and consultation within a contingency model of third party intervention. *Journal of Peace Research*, *28*(1), 29–42.
Franzese, J., Robert, J., & Hays, J. C. (2007). Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. *Political Analysis*, *15*(2), 140–164.
Gartzke, E., & Gleditsch, K. S. (2004). Why democracies may actually be less reliable allies. *American Journal of Political Science*, *48*(4), 775–795.
Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (p. 2007). Cambridge: Cambridge University Press.
Gibler, D. M. (2008). The costs of reneging: reputation and alliance formation. *Journal of Conflict Resolution*, *52*(3), 426–454.
Gleditsch, K. S., & Beardsley, K. (2004). Nosy neighbors: third-party actors in Central American conflicts. *The Journal of Conflict Resolution*, *48*(3), 379–402.
Greene, W. H. (2008). *Econometric analysis*. Upper Saddle River: Prentice Hall.
Hartley, K., & Sandler, T. (1999). NATO burden-sharing: past and future. *Journal of Peace Research*, *36*(6), 665–680.
Hashler, M., Gruen, B., & Hornik, K. (2009). A rules: mining association rules and frequent itemsets. R package version 0.9-6.
Haynie, S. L. (1992). Leadership and consensus on the U.S. Supreme Court. *The Journal of Politics*, *54*(4), 1158–1169.
Hix, S., Noury, A., & Roland, G. (2005). Power to the parties: cohesion and competition in the European Parliament, 1979–2001. *British Journal of Political Science*, *35*(02), 209–234.
Hjorth, J. U. (1993). *Computer intensive statistical methods*. Boca Raton: Chapman Hall/CRC.
Horne, J. S., & Garton, E. O. (2006). Likelihood cross-validation versus least squares cross-validation for choosing the smoothing parameter in kernel home-range analysis. *Journal of Wildlife Management*, *70*(3), 641–648.
Ireland, M. J., & Gartner, S. S. (2001). Time to fight: government type and conflict initiation in parliamentary systems. *The Journal of Conflict Resolution*, *45*(5), 547–568.
Jensen, D. D., & Cohen, P. R. (2000). Multiple comparisons in induction algorithms. *Machine Learning*, *38*(3), 309–338.
Johnson, T. R., Spriggs, J. F., & Wahlbeck, P. J. (2005). Passing and strategic voting on the U.S. Supreme Court. *Law & Society Review*, *39*(2), 349–377.
Johnson, T., Wahlbeck, P., & Spriggs, J. (2006). The influence of oral arguments on the U.S. Supreme Court. *American Political Science Review*, *100*(01), 99–113.
Kim, C.-H. (1991). Third-party participation in wars. *The Journal of Conflict Resolution*, *35*(4), 659–677.
King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: improving interpretation and presentation. *American Journal of Political Science*, *44*(2), 347–361.

Knafl, G. J., & Gray, M. (2007). Factor analysis model evaluation through likelihood cross-validation. *Statistical Methods in Medical Research*, *16*(2), 77–102.

Lebovic, J. H. (2004). Uniting for peace? Democracies and United Nations peace operations after the cold war. *The Journal of Conflict Resolution*, *48*(6), 910–936.

Leeds, B. A. (2005). Alliance treaty obligations and provisions (atop) codebook. *Houston: Rice University, Department of Political Science*.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3), 1142–1160.

Ostrom, E. (1998). A behavioral approach to the rational choice theory of collective action: presidential address, American Political Science Association, 1997. *The American Political Science Review*, *92*(1), 1–22.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151–1172.

Sandler, T. (2005). Collective versus unilateral responses to terrorism. *Public Choice*, *124*(1/2), 75–93.

Segal, J. A., & Spaeth, H. J. (2002). *The Supreme Court and the attitudinal model revisited*. Cambridge: Cambridge University Press.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, *9*(1), 63–72.

Solana, J. (1999). NATO's success in Kosovo. *Foreign Affairs*, *78*(6), 114–120.

Spriggs, I., James, F., & Hansford, T. G. (2001). Explaining the overruling of U.S. Supreme Court precedent. *The Journal of Politics*, *63*(4), 1091–1111.

Tucker, R. W., & Hendrickson, D. C. (2004). The sources of American legitimacy. *Foreign Affairs*, *83*(6), 18–32.

van der Laan, M. J., Dudiot, S., & Keles, S. (2004). Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular. Biology*, 3(1), Article 4.

Verba, S. (1961). Assumptions of rationality and non-rationality in models of the international system. *World Politics*, *14*(1), 93–117.

Ward, M. D., Siverson, R. M., & Cao, X. (2007). Disputes, democracies, and dependencies: a reexamination of the Kantian peace. *American Journal of Political Science*, *51*(3), 583–601.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25.

Wilkenfeld, J., Young, K., Asal, V., & Quinn, D. (2003). Mediating international crises: cross-national and experimental perspectives. *The Journal of Conflict Resolution*, *47*(3), 279–301.