Article

European Union Politics
2014, Vol. 15(1) 43–58

DOI: 10.1177/1465116513501908
eup.sagepub.com

Testing models of legislative decision-making with measurement error: The robust predictive power of bargaining models over procedural models

Justin Leinaweaver
Drury University, Springfield, MO, USA

Robert Thomson
University of Strathclyde, Glasgow, UK

Abstract

Previous studies found that models emphasising legislative procedures make less accurate predictions of decision outcomes in the EU than the compromise model, a computationally simple variant of the Nash Bargaining Solution. In this journal, Slapin (2014) argues that this and other findings may be the result of measurement error. While acknowledging the importance of measurement error, we disagree with several assumptions in Slapin's analysis, and show that his results are driven by an unrealistic assumption about how policy preferences are distributed among EU decision makers. We construct simulated data that more accurately reflect the distributions of policy preferences found in existing empirical evidence and suggested by theory, and demonstrate that measurement error is unlikely to have biased previous findings. If real-world decision-making took place according to the procedural model, then it would have made the most accurate predictions, even with data containing large amounts of measurement error. While this strengthens our confidence in previous studies' findings, we explain why we should not discard procedural models.

Keywords

Bargaining models, legislative decision-making, measurement error, procedural models

Corresponding author:
Robert Thomson, School of Government and Public Policy, University of Strathclyde, McCance Building, Level 4, Glasgow G1 1XQ, UK.
Email: robert.thomson@strath.ac.uk

Introduction

Procedural models and bargaining models offer a range of perspectives on how the European Union (EU) decides on legislative proposals. Procedural models analyse the EU's rules of procedure to identify the expected decision outcome based on a given distribution of policy preferences (e.g. Crombez, 1996; Steunenberg, 1994; Tsebelis, 1994; Tsebelis and Garrett, 2000). Bargaining models, by contrast, focus on the unwritten or informal rules of the political game (e.g. Achen, 2006a; Bueno de Mesquita and Stokmam, 1994). Procedural and bargaining models have been subjected to empirical scrutiny in a research project known as the DEU project (Thomson et al., 2006, 2012). This project focused on decisions on controversies during the post-proposal stage in the EU, after legislative proposals were introduced. A consistent finding was that procedural models generally made relatively inaccurate predictions of decision outcomes on these controversies, while the compromise model (CM), a computationally simple variant of the Nash Bargaining Solution (NBS), consistently made among the most accurate predictions (Achen, 2006b; Thomson, 2011).

To this point, the DEU project has not considered the possible effects of measurement error on the relative accuracy of models' predictions of decision outcomes.[1] Slapin (2014) argues that measurement error could drive the superior predictive accuracy of the CM over procedural models, as well as the findings of related analyses, which show that the Council is the most powerful of the institutions. To support his argument, Slapin creates a simulated world in which decisions are taken according to the logic of a procedural model. He then assumes that empirical research can only measure actors' policy preferences imperfectly. Slapin's simulations model the effects of different levels of measurement error on observed model performance in terms of predictive accuracy. His findings indicate that as the amount of measurement error increases, the procedural model makes less accurate predictions of decision outcomes than the bargaining models, even though decisions in this simulated world are taken according to the procedural model. The implication, Slapin argues, is that the findings of the DEU project could be driven by measurement error, and therefore we should not conclude that procedural models actually make less accurate predictions than the CM just because the empirical results from the DEU project indicate that they do.

We offer a two-part response that engages with Slapin's analysis and conclusions. First, we acknowledge the importance of gauging the effect of measurement error on the key findings of the DEU project, but disagree with the main assumptions in Slapin's analysis and simulations. With more realistic assumptions about the types of measurement error contained in the DEU data, we conclude that the DEU project's findings are robust. If decision-making really did take place in line with the procedural model, then the procedural model would have made more accurate predictions than bargaining models, even if the data contained large amounts of measurement error. We focus on Slapin's critique of the DEU project's findings regarding procedural and bargaining models' relative predictive accuracy. However, our response is just as relevant to his criticisms of other findings using the DEU dataset, since these criticisms also contend that measurement error is what

drives these findings. Second, we argue that procedural models should not be discarded based on the fact that they on average make less accurate predictions than some other models.

A realistic assessment of the effect of measurement error on predictive accuracy

To understand the types of measurement error we are likely to encounter with the DEU data and design an appropriate analysis of the effects of measurement error, we must consider what these data look like. The top of Figure 1 gives an example of a typical controversial issue. This controversy was raised by a proposal concerning the organisation of the support scheme for olive oil producers and the categorisation and labelling of various oils. The controversial issue was the question of what to do with mixtures of olive oils and vegetable oils. At the left of the policy scale used to represent this controversy, position zero refers to the then current state of affairs that would have prevailed in the absence of the new regulation. Under this scenario, mixed oils would not be subject to any new requirements. At the right of the policy scale, position 100 refers to the most extreme position taken by the olive oil producing countries (Greece, Italy, Portugal and Spain). They wanted to ban such mixtures. The Commission and most member states agreed that the current situation needed to be changed, in that clearer labelling was required. A key

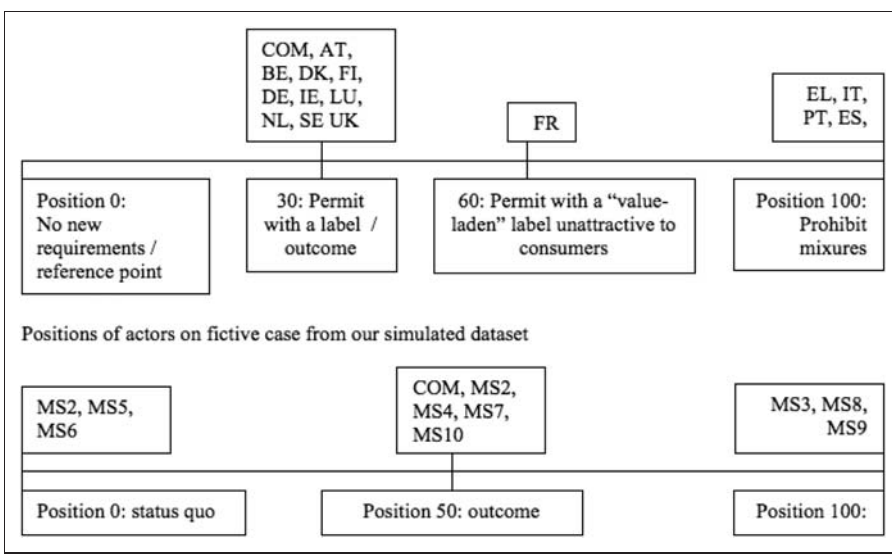

Figure 1. A real case and a fictive case from our simulated dataset. The real case refers to a controversy raised by the proposed regulation regarding the aid scheme and quality strategy for olive oil (CNS/2000/358).

informant we interviewed placed this outcome on position 30, to reflect his judgement that this alternative was politically closer to having no new requirements than to prohibiting mixtures. The French permanent representation favoured a 'value-laden' label being attached to mixtures, which would have indicated their inferior quality and would therefore have been unattractive to consumers (placed at position 60 by the key informant). The Southern member states' policy positions are clearly in line with their economic interests and culinary tastes. In this case, the outcome was in line with the Commission's legislative proposal.[2] The olive oil controversy is typical of the data contained in the DEU dataset, in that actors are clustered around a limited number of policy positions. On average, there are 3.31 distinct policy positions (standard deviation 1.55) across the 331 controversial issues contained in the latest version of the DEU data.

With this example in mind, we discuss our disagreement with four of the assumptions inherent in Slapin's analysis. The first of Slapin's assumptions with which we disagree is that tests of linear fit are appropriate for assessing the relative predictive accuracy of competing models. He regresses actual decision outcomes on predicted outcomes in a model with no constant and interprets the size of the coefficients as a measure of predictive accuracy. This is a problematic method that can lead to erroneous conclusions for several reasons. When we assess predictive accuracy, we are concerned with how close a model's predictions are to the actual outcomes on average, not whether there is a linear fit between the model's predictions and actual outcomes. Slapin's regressions may identify a model that makes predictions further from actual outcomes as the 'winner'. Consider the following scenario depicted in Figure 2. Assume that actual outcomes are drawn from a normal distribution with a mean of 50 and a standard deviation of 10. Suppose that the predictions of model 1 are also drawn from a normal distribution with a mean of 50 and standard deviation of 10. The predictions of model 1 are fairly close to the decision outcomes (model 1 had a mean absolute error of 11.07 in the 500 draws we took). Suppose that model 2 is consistently 20 points to the left of each decision outcome. Although model 1's predictions are more accurate than those of model 2, the regression with no constant that Slapin applies would suggest the opposite. The coefficient associated with model 1 would be less than a third of that associated with model 2 (.28, with a standard error of .01, compared to 1.18, with a standard error of .01, in the simulation we ran, depicted in Figure 2). Although the precise distributions of decision outcomes and predictions depicted in Figure 2 are unlikely to occur in any real or simulated dataset, similar distributions can occur. Indeed, two studies demonstrate that empirical tests based on linearity, such as regression, rather than closeness, such as absolute distance between predictions and outcomes, lead to substantively different results (Achen, 2006b: 276, 285; Dijkstra et al., 2008: 429).

Another reason why regression is problematic is that the predictions of different models are often highly correlated with each other. The dependent variables in Slapin's regressions are the decision outcomes and the independent variables are the predictions of decision outcomes made by different models. Since the models

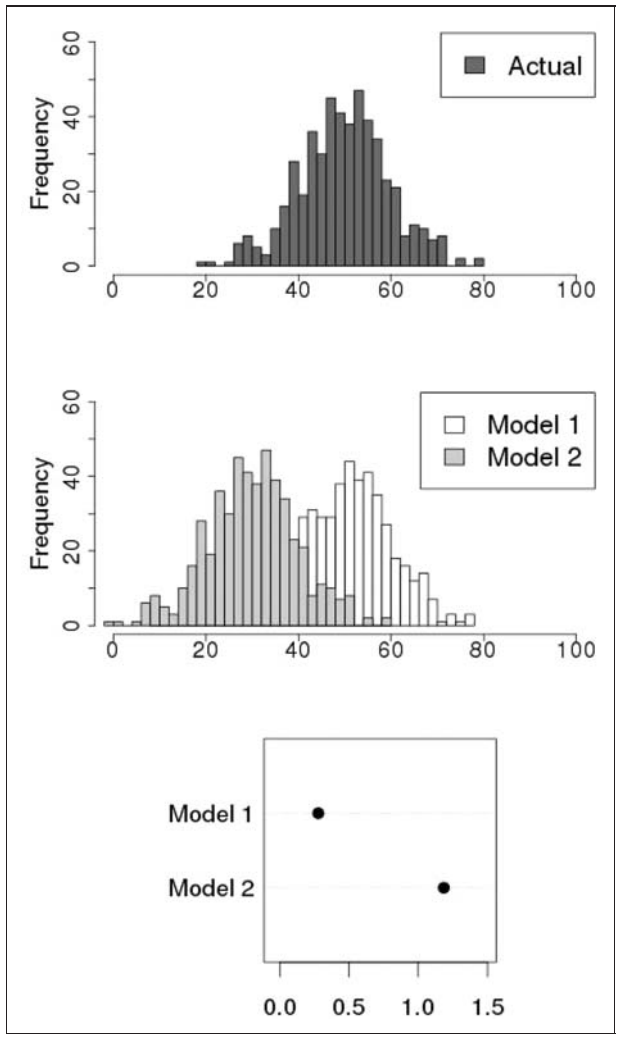

Figure 2. Why regression coefficients are problematic measures of relative predictive accuracy. Top: distribution of actual decision outcomes; middle: distributions of relatively accurate (Model 1 in light grey) and relatively inaccurate (Model 2 in dark grey) predictions of these decision outcomes; bottom: visualisation of regression coefficients showing that the coefficient for the accurate model is significantly smaller than that of the inaccurate model.

derive their predictions from the same fundamental components of the bargaining situation (mainly actors' policy positions on policy scales), it is hardly surprising that their predictions are often correlated. Finally, regression is inappropriate when the direction of scaling of many issues is arbitrary. While the DEU project followed the convention of placing positions that corresponded with 'more EU'

closer to position 100, over a third of the issues (115 of 331) could not be coded in this way (Thomson, 2011: 69). Although actors took distinct positions on these 115 issues, it was not possible to say that some actors were calling for 'more EU' than others. The direction in which these issues are scaled (i.e. which positions are placed at which endpoints, 0 or 100, of the policy scales) is arbitrary, but does affect the regression results. We prefer to follow the convention of taking the mean of the absolute distances between a model's predictions of decision outcomes and actual outcomes as our criterion of assessment.[3]

Second, we disagree with Slapin's assumption that a two-actor framework, consisting of only the agenda setter and pivotal actor, is a credible analysis of the impact of measurement error on models' predictive accuracy. The key point is that the pivotal actor is identified in relation to the positions of all other committee members who must approve the agenda setter's proposal for that proposal to pass. The analysis should therefore consider the possibility that errors are made in the estimation of those other committee members' policy positions, not just the policy positions of the pivotal actors. Tsebelis and Garrett (2000) analyse a hypothetical committee of seven members to explicate their procedural model. In the following analyses we consider a committee of 10 members, which allows us to explore more realistic scenarios regarding measurement error.

The third assumption with which we disagree is that error is just as likely in measuring the location of all actors' positions, the status quo and the outcome. In the DEU project, error is more likely to be found in member states' policy positions than the positions of the Commission, status quo or outcome. The DEU dataset is based on semi-structured interviews with key informants, who were usually involved in the negotiations themselves. These individuals were well aware of the details of the negotiations, the interests of the various actors and the available (often restricted) documentation. Outcomes are recorded in the contents of laws and the status quo positions (or reference points as they were referred to in the DEU project) often refer to the existing legally defined situations. While the European Commission issues a legislative proposal that reflects the outcomes it at least claims to favour, no such documents are available for each member state. Therefore, key informants often needed to rely on what they heard representatives of other member states saying in Council meetings when reporting those member states' positions. It therefore seems far less likely that key informants reported and researchers did not discover inaccuracies in estimates of the positions of the Commission, reference points or outcome positions than member states' positions. In the example of the olive oil controversy, we were able to confirm the positions of the Commission, the outcome and the reference point from the legislative proposal (including the background statement contained in the proposal) and the final legislative act. Regarding the member states' positions, it seems unlikely that the data do not reflect the positions actually favoured. As illustrated by the olive oil controversy, states' positions usually reflect their interests. Nonetheless, we do not deny the possibility of measurement errors with respect to member states' positions and indeed consider it in our own simulations below.

The fourth and final assumption with which we disagree is that actors' policy positions are uniformly distributed. As will become clear below, this is the key difference between us and Slapin. The olive oil controversy, which is typical of the data contained in the DEU dataset with respect to its clustering, markedly differs from Slapin's simulated data. In the real world, actors are clustered around a limited number of policy positions, while in Slapin's simulated world, they are uniformly distributed across the policy scale. Slapin assumes that each actor has an equal probability of taking any of 60 policy positions on each policy scale ('A's ideal points, x_{Aj}, are drawn from $S_{Aj} = \{0, 1, 2, 3, \ldots, 60\}$, while B's ideal points, x_{Bj}, are drawn from $S_{Bj} = \{40, 41, 42, 43, \ldots, 100\}$' (2014: 24–42)). The uniform distribution is more than just an analytical convenience; it is the assumption that drives his main finding.

In many contexts the distribution of actors' policy positions depends on whether ideological dimensions or specific policy issues are at stake. Slapin's references to 'ideology' imply he is concerned with the former, while the DEU project is concerned with the latter. We suggest that ideology is of limited relevance to position taking in the Council of Ministers. Ideological dimensions refer to choices between abstract principles that might guide a range of relevant decisions. The most commonly cited of these is the left–right dimension, which often refers to the degree of state intervention in markets and social policies. By contrast, the DEU project's specific policy issues refer to concrete choices among different policy alternatives, as the example in Figure 1 makes clear. While there is evidence that partisan ideology is relevant to position taking in the European Parliament, there is no evidence that it is generally relevant to position taking on specific policy issues in the Council. So when member state representatives take issues on the olive oil controversy, they tend to be thinking of their states' interests in the olive oil sector, rather than their ideological preference for a certain degree of European integration or a certain degree of state intervention into free markets.[4]

Our analyses will make clear that it is this fourth disagreement that leads to a different conclusion regarding the effect of measurement error on models' predictive accuracy. We begin by replicating Slapin's analysis, but with two differences, referring to the first two assumptions mentioned above. We use the mean average error as the criterion for assessing models' performance and use a framework consisting of 10 committee members and an agenda-setter, which correspond to 10 hypothetical member states and the Commission. We begin by creating a world with no measurement error, in which players decide on 150 independent policy issues or policy scales. Each scale ranges from zero to 100, where larger numbers may be thought of as pro-integration positions. In this world, the decision-making process is defined by the procedural model, referred to as the Romer-Rosenthal, or R-R model. As in Slapin's analysis, we set the status quo on each issue to position zero. Each committee member's position is drawn randomly from a uniform distribution of whole numbers ranging from zero to 60. The agenda setter's position is drawn randomly from a uniform distribution of whole numbers ranging from 40 to 100. For the Commission's policy proposal to pass, it must receive the support of

seven of the 10 member states, which approximates the level of support needed for a bill to pass under qualified majority voting. We calculate the R-R model in the same way as Slapin, which means that the decision outcome on an issue is equal to:

$$2 \times |p - s| + s \quad \text{if } a > 2 \times |p - s| + s$$

where p, a and s are the positions of the pivot, agenda setter and status quo, respectively. The outcome is equal to the position of the agenda setter, a, if the above condition is not satisfied, which means that the pivotal actor prefers the agenda setter's position to the status quo.

Following Slapin's analysis, we then add measurement error to our observations of this simulated world by adding normally distributed measurement error to the position of each of the 10 committee members, the agenda setter, the actual decision outcome (as defined by the R-R model) and the status quo. We re-scale these error-laden estimates of the positions to ensure that all values remain on our 100-point scale. We then apply the R-R model, NBS and CM to identify the outcomes they predict based on these error-laden estimates. We also define the NBS and CM in the same way as Slapin. Recall that the CM is simply the mean average of actors' positions. Since the R-R model is the true model in our simulated world, this model should make the most accurate predictions of decision outcomes in our observed data unless measurement error is distorting the results. Figure 3 summarises the findings of this first simulation. At low levels of measurement error (with a mean of zero and standard deviation of two), across 1000 iterations of the

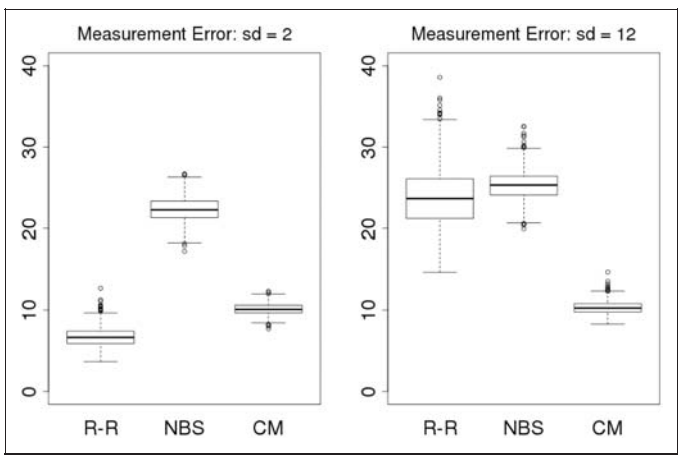

Figure 3. Model performance is sensitive to measurement error if policy positions are uniformly distributed. Replication based on Slapin's (2014) analysis, but with a mean average error as the criterion for model assessment and a 10-actor framework. Boxplots show distributions of mean average errors of the predictions of the Romer-Rosenthal model (R-R), Nash Bargaining Solution (NBS) and compromise model (CM) across 1000 iterations on 150 issues.

simulation, the R-R model makes the most accurate predictions. The R-R model has a mean absolute error of on average 6.7 scale points away from the actual outcomes, compared to the CM with an error of 10.1 and the NBS with an error of 22.3. In line with Slapin's main finding, when the measurement error is considerably larger (with a mean of zero standard deviation of twelve) the CM appears to make the most accurate predictions with a mean error of 10.3, compared to the R-R at 23.8 and the NBS at 25.3.

We now turn to what we consider to be a more realistic assessment of the effects of measurement error on model performance, by taking into account a key feature of the DEU data, namely that actors' policy positions are clustered on a limited number of points on each policy scale. Our simulations focus on the two main types of possible measurement error: misallocation of actors into clusters and misplacement of clusters relative to each other. We follow the same, 10-actor setup as above with 150 issues or policy scales. This time, however, actors' policy positions are clustered on three points (0, 50 and 100 of each policy scale), rather than being distributed uniformly across (large parts of) the policy scale. Each of the 10 committee members is randomly allocated to one of the clusters (positions 0, 50 or 100) on each of the 150 issues. The agenda setter is then allocated randomly to either position 50 or 100, to reflect its preference for deeper integration. The status quo is again set to position zero. As above, the R-R model defines the true decision-making process, which means that for the agenda setter's policy proposal to pass, it must receive the support of seven committee members. In the scenario depicted in the bottom of Figure 1, the pivotal committee member or member state is at position 50, since there are only three member states on position zero (the status quo). In that example, since the location of the Commission and the Council pivot are the same, this is also the actual outcome.

To simulate the effect of realistic measurement error on our observations of models' predictive accuracy, we examine different combinations of our two types of measurement errors:

a. one of the 10 committee members is allocated incorrectly to one of the other two clusters to which it in fact does not belong;
b. two of the 10 committee members are allocated incorrectly to clusters to which they in fact do not belong;
c. an error is made in the relative positioning of the three clusters. In this first variant of mispositioning, the middle cluster, which actually has position 50, is replaced by with a random draw from a normal distribution with a mean of 50 and a standard deviation of 10. All of the actors (and the true outcome and/or agenda setter if they happen to be located at position 50) are moved to this new incorrect position.
d. a large error is made in the relative positioning of the three clusters. In this second variant of mispositioning, the middle cluster is replaced as in variant c, but by a random draw from a normal distribution with a mean of 50 and a standard deviation of 20.

e. a combined error of two of the 10 committee members being misallocated (variant b) and a large error in the relative positioning of the three clusters (variant d).
f. an error is made in allocating either the status quo, outcome, agenda setter or one of the 10 committee members to a cluster to which it does not belong. This final robustness test allows us to examine whether the differences in our findings are due to our disagreement with Slapin's assumption that errors are equally likely in estimating all positions.

For each of the above six variants of measurement error, we ran 1000 iterations on each of the 150 issues in our simulated world. In each iteration we created a new true simulated world by assigning actors' positions randomly, and then identifying the pivot and the true outcome according to the R-R model. Within each of the 1000 iterations we then added the relevant type of measurement error, and calculated the observed prediction and observed absolute error of each of the three models. We then calculated the average absolute distance between each model's predictions and the observed outcomes across those 1000 simulated iterations for each issue. In the following, we examine the distributions of the average absolute errors of the three models across the 150 issues in our simulated world in the six different variants of measurement error.

The findings depicted in Figure 4 indicate that even at high levels of measurement error, the R-R model's predictions are far more accurate than those of the NBS or CM. If there were no measurement errors, the R-R model's predictions would make no errors in predicting decision outcomes, because this is how decisions are taken in our simulated world.[5] When we misallocated one of the 10 committee members to a cluster to which it did not belong (variant a: 10% misallocation), the R-R model made less accurate predictions of decision outcomes than it did in the absence of measurement error, but its predictions of decision outcomes were still far more accurate on average than those of the NBS or the CM (top left of Figure 4). We observed a similar pattern when we misallocated two of the 10 member states to clusters to which they did not belong (variant b: 20% misallocation; top right of Figure 4). Mispositioning the clusters relative to each other had smaller effects on the relative predictive accuracy of the three models (variants c and d: mispositioning with standard deviations of 10 and 20 points on the policy scale; middle left and right of Figure 4). The bottom left panel of Figure 4 depicts what happened to the prediction errors when we combined the worst two types of errors: misallocating two of the 10 member states and mispositioning the clusters relative to each other by an average error of 20 policy scale points (variant e above). Even then, the R-R model still made more accurate predictions of decision outcomes than the NBS or CM. Finally, in variant f we relaxed our assumption that the status quo positions, outcomes and agenda setter's positions are accurately measured. Even then, the R-R model outperformed the other two models in terms of predictive accuracy. Together, these analyses show that while we disagree with four assumptions in Slapin's analysis, it is our disagreement with

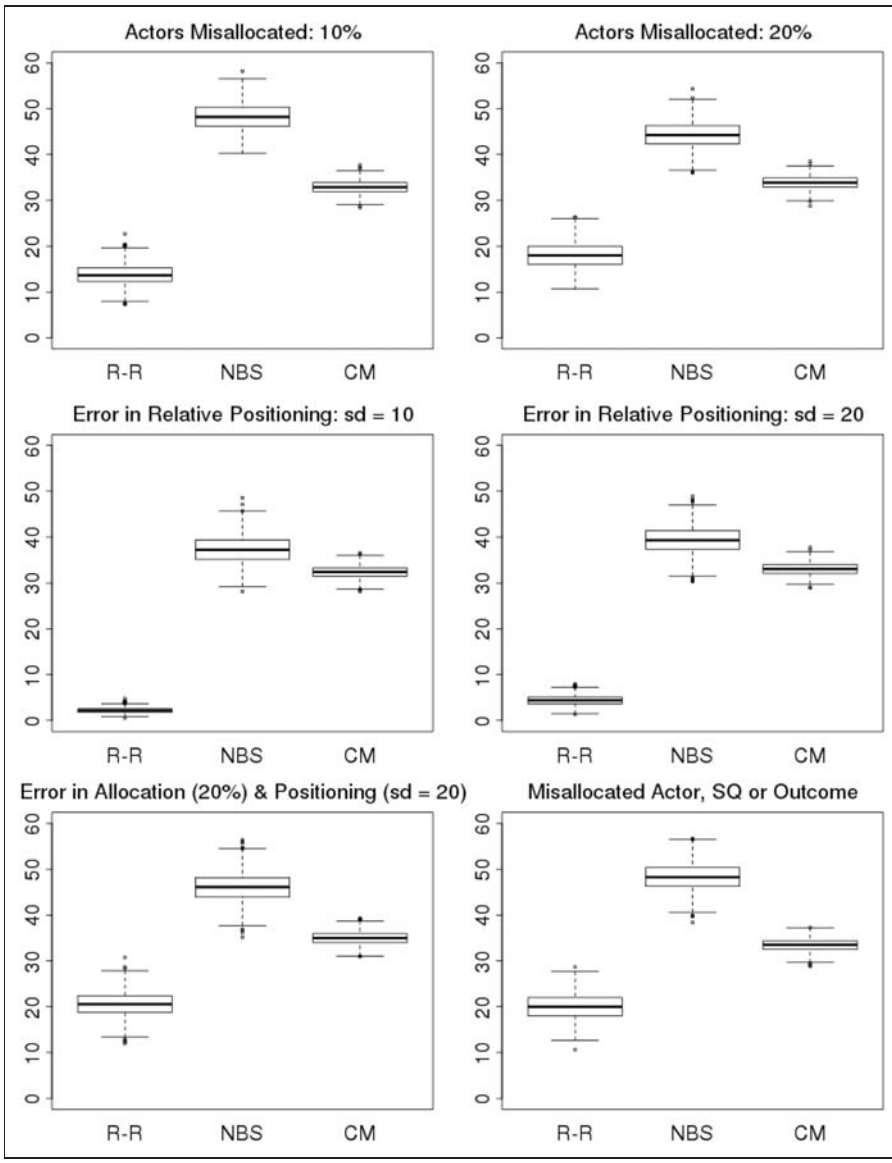

Figure 4. Model performance is robust to measurement error if policy positions are clustered. Distribution of mean average errors of the predictions of the Romer-Rosenthal model (R-R), Nash Bargaining Solution (NBS) and compromise model (CM) across 1000 iterations on 150 issues.

the assumption of uniformly distributed policy positions that leads us to a fundamentally different conclusion regarding the likely effect of measurement error.

Our critics may argue that our simulated world does not reflect the unobserved, and perhaps unobservable world in which actors' positions are uniformly distributed. In response, we note that we are guided by both the best available evidence on actors' preferences on specific issues in EU decision-making and well-established theory when constructing our simulated world. The evidence from the DEU project indicates that actors' positions are clustered around an average of three (3.31) distinct positions on each policy scale. Moreover, liberal international relations theory leads us to expect such clustering in states' positions. According to this theory, states' preferences are defined by the balance of domestic interests, as well as significant patterns of international interdependence (Moravcsik, 1997: 523). EU member states are certainly embedded in significant patterns of interdependence, and we would therefore expect their preferences to be affected by one another, which is consistent with the observed clustering of positions.

Insights from procedural models

One of the main findings from the DEU project is that procedural models make relatively inaccurate predictions of decision outcomes on controversies raised by legislative proposals in the EU (Achen, 2006b; Thomson, 2011). Other models that place great emphasis on the disagreement outcome, including some bargaining models, also make relatively inaccurate predictions. By contrast, the CM consistently ranks among the best predictors of decision outcomes. Despite the CM's computational simplicity (it is simply an average or weighted average of policy positions), it has important theoretical properties. As the disagreement outcome becomes less and less desirable from the perspective of all of the actors, the NBS approximates, and at its limit is equal to, the CM (Achen, 2006a). Slapin cautions that on the basis of models' relative predictive power 'we would not want to conclude that the institutional models are wrong' (2014: 24–42), and by 'institutional models' he is referring to procedural models. We agree with Slapin here, but not for the reasons he gives. In this section we briefly set out why procedural models should not be discarded.

First, studies of controversies that persist and arise after legislative proposals are introduced focus on situations in which procedural models are unlikely to make accurate predictions. By design, the DEU project selected legislative proposals that raised at least some controversy after they were introduced and asked how the EU decided on such controversies. Many procedural models would not expect to encounter much explicit controversy at all after legislative proposals have been introduced. In the standard procedural model, the agenda setter cleverly selects the policy alternative that will be supported by the pivotal actor in the relevant committee; the agenda setters' proposals should then pass without amendment. The fact that procedural models predict decision outcomes poorly in situations they do not expect to encounter is not grounds on which to discard them. Despite the fact that important controversies often do persist and arise after the

introduction of legislative proposals, it could be argued that a large proportion of legislative proposals are adopted with relatively little fuss, as the standard procedural model leads us to expect. Moreover, even when a legislative proposal does raise important controversies, these controversies often refer to a relatively small number of provisions. So perhaps if we were to define our universe of relevant observations differently, as the universe of policy changes introduced by the Commission rather than the universe of controversial issues raised by policy changes introduced by the Commission, procedural models would not fare as poorly in terms of the accuracy of their predictions of decision outcomes.[6]

Second, procedural models offer insights into the post-proposal stage of decision-making that have yet to be fully explored. In the largest comparative test of predictive accuracy to date, which focused on 233 issues, one procedural model made more accurate predictions than the CM on 91 issues (39%) and less accurate predictions on 142 issues (61%; Thomson, 2011: 183). Until now research has focused on generalisations about the model that makes the most accurate predictions across all issues. The obvious next step is to devote more attention to the conditions under which different models are more applicable. One possible condition concerns the relevance of the disagreement outcome. It could be that procedural models are more applicable when the disagreement outcome is less unlikely. Challenges regarding conceptualisation and measurement need to be overcome to measure the disagreement outcome adequately, because failure to agree means not only that the controversial issue in question is unresolved, but also that the parts of the proposal on which there is a consensus are lost and perhaps also that damage is done to the relationships among the actors (Achen, 2006a: 101). However, such effort would be worthwhile to advance our understanding of these important political processes.

Even when cooperative bargaining models make more accurate predictions of decision outcomes during the post-proposal stage than do procedural models, procedural rules may nonetheless be relevant. The consensual and inclusive nature of decision-making in the EU, where decisions in the Council of the EU are usually passed by more than the bare qualified majority of member states, is supported by both variation in the alignments of actors found across different specific controversies as well as by the procedural rules (Schneider et al., 2006: 305). Because member states find themselves in different coalitions on different issues, each member state knows that even if it is not at present part of a potentially losing minority that could be outvoted according to the procedural model, it could be on another occasion. The implicit threat of being outvoted in the future compels member states to take into account minority opinions. It is therefore the procedural rules that underpin the norm of consensus seeking among member states. We hope that future research will specify this rule-based mechanism behind the norm of consensus in a more testable form and compare it with other possible mechanisms.

Third, procedural models should not be discarded on the basis of the findings from the DEU project, because they offer insights to constitutional analysts and engineers when exploring the possible effects of changes to decision rules.

The Treaty of Lisbon is the latest outcome in a long series of negotiations in which EU decision makers have considered alternative rules of procedure. For instance, Tsebelis (2012) examines the effects of changing the triple majority rule governing Council voting in the Nice Treaty with the double majority rule in the Lisbon Treaty, and demonstrates that the double majority rule should make it easier to pass legislation. Using procedural models, analysts provide valuable indications of the maximum reasonably possible effects of such rule changes. What the actual effects of the rule changes will be depends on how actors interpret those rules and the coalitions that will form in the future. We might take past experience as a guide to the future and note that the clustering of actors' positions could minimise the actual effect of rule changes. Groups of member states that are blocking minorities according to the Nice rules are usually also blocking minorities according to the Lisbon rules (Thomson, 2011: 184–185). This does not, however, detract from the importance of knowing what the maximum reasonably possible effects of rule changes could be, or from the need to collect new data on how the EU decides on controversial matters after the new rules have come into effect.

Conclusions

We analysed the effects of measurement error on the accuracy of different models' predictions of decision outcomes. We did so in relation to the data and key findings of the DEU project (Thomson et al., 2006, 2012), which examined specific controversies raised by legislative proposals in the EU and the predictive accuracy of procedural and bargaining models. The key finding of our simulation-based analysis is that if the decision-making process took place according to a procedural model, then the empirical findings would reflect this, even if the data on actors' positions contained high levels of measurement error. Research conducted in the DEU project found that procedural models make significantly less accurate predictions than other models, including the CM. The simulation-based analysis presented here therefore strengthens our confidence in this key finding from the DEU project. The superior predictive accuracy of the CM over the procedural model reflects the reality of the decision-making process, not measurement error. Our simulation-based analysis takes into account the clustering of actors' policy positions into discrete groups, while Slapin's simulations that inspired our analysis do not. While we disagree with several assumptions in Slapin's analysis, we show that it is our disagreement on the most appropriate assumption regarding the distribution of actors' policy positions that drives our different conclusions. For us, the clustering of actors' policy positions is a key aspect of decision-making that we would expect based on existing empirical research and theory, and one that should be incorporated into analyses of the impact of measurement error.

We also enumerated several reasons why procedural models should not be discarded despite the fact that their predictions of decision outcomes are less accurate than those of other models. Procedural models have enriched our

understanding of decision-making in ways that go beyond the empirical confines of the DEU project (Crombez, 1996; Finke et al., 2012; Steunenberg, 1994; Tsebelis, 1994; Tsebelis and Garrett, 2000). The DEU project focused squarely on how the EU decides on controversies that persist and arise after legislative proposals have been introduced. While this is an important stage of the decision-making process, the standard procedural model would not expect the agenda setter's proposal to be amended, and perhaps also not the subject of much debate among decision makers at all. In other words, it could be argued that the DEU project focuses on cases in which the procedural model has failed. Nonetheless, we believe that procedural models could still have much to teach us about what happens during this post-proposal stage, partly because procedural models make relatively accurate predictions on a substantial minority of issues. Future research should also consider alternative research designs for testing procedural and bargaining models in ways that move beyond the focus on explicit controversies during the post-proposal stage.

Notes

1. The DEU project extensively assessed the reliability and validity of the estimates gathered (see appendix II of Thomson et al., 2006), although did not consider the implications of possible measurement error for the relative accuracy of models' predictions.
2. Council Regulation (EC) No 1513/2001. The ninth recital states that blends should be 'identified in a specific way without, however, detracting from the qualities of the type of oil concerned, which are appreciated by a large sector of the market.'
3. We do not believe the linear-fit approach has any noteworthy advantages over the closeness approach in this context. The closeness approach can also be used to test whether models' errors differ statistically significantly from each other, for instance using a non-parametric sign test (Achen, 2006b: 290), which is appropriate in many datasets in which the issues are not entirely independent observations. The arbitrary direction of scaling of many issues in the DEU dataset renders Slapin's empirical analysis in the second part of his paper problematic. Space precludes us from a more extensive discussion of Slapin's empirical analysis, but we note that the use of imputed positions for actors that key informants reported were indifferent is another analytical decision with which we disagree.
4. Our argument is that member states' policy positions on policy issues are generally clustered, not that they are inherently categorical or ordinal in nature. Indeed, key informants find it possible to place groups of actors at different points on each policy scale to reflect the political distances between the policy alternatives they favour.
5. In a hypothetical world with no measurement error, while the R-R model has an error of zero, the NBS has an error of 36.68 (standard deviation (s.d.) 3.01) and the CM has an error of 31.99 (s.d. 1.37) in our simulated world with 150 issues over 1000 iterations.
6. Our argument here is similar to Achen and Snidal's (1989) discussion of the case study evidence that contradicts rational deterrence theory. They point out that these case studies generally refer to violent conflict, which by definition are cases in which rational deterrence theory has failed, while ignoring the multitude of seemingly uninteresting cases where the theory successfully explained why conflict was averted.

References

Achen CH (2006a) Institutional realism and bargaining models. In: Thomson R, Stokman FN, Achen CH and König T (eds) *The European Union Decides*. Cambridge: Cambridge University Press, pp. 86–123.

Achen CH (2006b) Evaluating political decision-making models. In: Thomson R, Stokman FN, Achen CH and König T (eds) *The European Union Decides*. Cambridge: Cambridge University Press, pp. 264–298.

Achen CH and Snidal D (1989) Rational deterrence theory and comparative case studies. *World Politics* 41(2): 143–169.

Bueno de Mesquita B and Stokman FN (1994) *European Community Decision Making: Models, Applications and Comparisons*. New Haven: Yale University Press.

Crombez C (1996) Legislative procedures in the European community. *British Journal of Political Science* 26(2): 199–228.

Dijkstra J, Van Assen MALM and Stokman FN (2008) Outcomes of collective decisions with externalities predicted. *Journal of Theoretical Politics* 20(4): 415–441.

Finke D, König T, Proksch SO, et al (2012) *Reforming the European Union: Realizing the Impossible*. Princeton: Princeton University Press.

Moravcsik A (1997) Taking preferences seriously: A liberal theory of international politics. *International Organization* 51(4): 513–553.

Schneider G, Steunenberg B and Widgrén M (2006) Evidence with insight: What models contribute to EU research. In: Thomson R, Stokman FN, Achen CH and König T (eds) *The European Union Decides*. Cambridge: Cambridge University Press, pp. 299–316.

Slapin JB (2014) Measurement, model testing and legislative influence in the European Union. *European Union Politics* 15(1): 24–42.

Steunenberg B (1994) Decision-making under different institutional arrangements: Legislation by the European community. *Journal of Theoretical and Institutional Economics* 150(4): 642–669.

Thomson R (2011) *Resolving Controversy in the European Union*. Cambridge: Cambridge University Press.

Thomson R, Arregui J, Leuffen D, et al (2012) A new dataset on decision-making in the European Union before and after the 2004 and 2007 enlargements (DEUII). *Journal of European Public Policy* 19(4): 604–622.

Thomson R, Stokman FN, Achen CH, et al (2006) *The European Union Decides*. Cambridge: Cambridge University Press.

Tsebelis G (1994) The power of the European Parliament as a conditional agenda setter. *American Political Science Review* 88(1): 128–142.

Tsebelis G (2012) From the European convention to the Lisbon Agreement and beyond: A veto player analysis. In: Finke D, König T, Proksch SO and Tsebelis G (eds) *Reforming the European Union: Realizing the Impossible*. Princeton: Princeton University Press, pp. 28–61.

Tsebelis G and Garrett G (2000) Legislative politics in the EU. *European Union Politics* 1(1): 9–36.