Evaluating Political Decisionmaking Models

(Chapter 10 of *The EU Decides*, forthcoming,
Cambridge University Press, 2005)

Christopher H. Achen
Department of Politics
Princeton University
Princeton, NJ 08544

April 16, 2005
How Well Do the Models Forecast?¹

Stocks have reached what looks like a permanently high plateau.

– Irving Fisher, prominent professor of mathematical economics
at Yale University, Oct. 17, 1929

The previous chapters in this book have elaborated many different models of political decision making in the European Union. To make their forecasts, some of these models focus on the incentives created by EU legal regulations or decision making rules. Other models emphasize the power of bargaining in political decision making. Still others start from log-rolling, coalitions, or the spatial theory of voting.

The aim of this book is to set out all these models of EU decision making, and then to evaluate how well the models predict actual decisions. Most of the discussion is quantitative. Yet in important respects, we mean to integrate prior case studies, formal theory, and statistical methods. For example, each of the modelling approaches represented in this book builds on one or more central aspects of political life known from dozens of skillful case studies of political decisions. All competent model building depends on careful qualitative research in which explanatory factors are identified and tentative empirical generalizations are formulated. No model is worthwhile if, like some formal theorizing, it applies to nothing in particular. Case studies have generated most of the interesting hypotheses in political science. They are the essential foundation for most model building.

Moreover, case studies play a crucial role in evaluating theory. Indeed, the entire dataset used in this book consists of short case studies, which are
then coded for quantitative analysis. Nearly all the chapters discuss one or more of these cases and demonstrate that the models under discussion succeed in explaining some aspects of the outcome. Failures are also discussed, with the aim of improving future theory. Thus for theory testing and criticism, too, we treat case studies as indispensable.

Yet case studies alone, like formal theories alone, have limitations. By their nature, case studies derive their conclusions from short periods of political history, often filled with intense conflict among vivid personalities. Analysts struggle to separate suggestive patterns from irrelevant distracting details. Even if they are successful in doing so, the findings, however insightful, are of unknown generality.

For all these reasons, even the best case study findings achieve greater precision and theoretical fecundity when they can be accounted for, at least in part, by a formal model (Achen and Snidal 1989). In fact, purely at the theoretical level, formulating models with precise predictions has sharpened our understanding of the empirical generalizations deriving from case studies and strengthened our conceptual grasp of EU decision making, as previous chapters in this book have demonstrated. The exactitude of the models’ formulations enables researchers to pinpoint our explanatory weaknesses and improve our knowledge in ways that were wholly unmatched and unforeseen in the days of purely verbal treatments of political decision making. In turn, this knowledge feeds back into case study analyses, sharpening their interpretations and adding power to their inferences. Thus the empirical accuracy of historical study and the theoretical precision of formal models are neither grand opposing explanatory frameworks nor minor differences of
technique. Instead, both are part of an integrated social science research methodology that adds to our comprehension of both theory and individual cases.

Lastly, statistical methods, too, are needed. For case studies plus theoretical sophistication alone do not suffice: Predictive accuracy across multiple cases matters, too. The interplay between models and case studies is not meant merely to summarize interesting history or to play out intriguing concepts, but rather to capture the precise workings of some central aspects of reality. Hence we need to know how well a model tracks the details of new political decisions, not just those it was constructed to explain.

Models reproduce only part of reality, and so all models sometimes fail. Yet some fail less than others. With a plethora of models in hand, then, we need comparative model evaluation. Here statistical methodology enters.

Thus having developed a range of models in previous chapters, each of them drawing on prior case studies, we arrive at the statistical evaluation of the models. How good are these models? How much of reality do they capture? And how well do they perform relative to very simple forecasting rules that require much less theory? Answers to these questions are obviously important for formal theorizing about the EU.

We believe, however, that the findings of this book are just as important for the case study tradition. Quantitative and qualitative methods relate to each in the manner of optical and radio telescopes: Sometimes they see the same things differently, and sometimes one sees what is invisible to the other. But their conclusions are mutually beneficial and reinforcing, making their joint value much higher than their individual parts. In the same way,
we hope that the results of our comparative model evaluations will inspire not only formal theorists, but also case study specialists and statistical empiricists, to build on our work. The best broad-gauge understandings of the EU will emerge from that dialog.

Forecasts and Baselines

No theory copes with all the peculiarities of politics. Proposal COM/2000/604 in our dataset, for example, contained three issues dealing with sugar quotas and storage subsidies, with two of the issues dichotomous (yes/no). On these two, the Commission was opposed by nearly everyone else, and some of the other actors felt strongly. Thus most of the models in this book foresee very probable defeat for the Commission.

In fact, however, the Commission got its way on both issues. Finland received an exemption from one clause; Italy, Portugal, and Spain were excused from another. These sidepayments quieted several of the Commission’s opponents, but they are not “issues” in the usual sense and so are not part of the dataset. Moreover, some unusual aspects of the decisions meant that the Commission could impose its will if the member states failed to agree. Thus the standard Shapley–Shubik values fail to represent the Commission’s real power in this particular case. In short, the opponents were much less concerned than they appear to be in our codings, and the Commission was far more weighty.

The result is a substantial prediction error for most of the models. The culprits are the unavoidable limitations of event coding and the daunting
complexity of the political world. Some errors of this kind are inevitable. Models and theories are meant to give insight into what usually happens. They are not meant to replace detailed descriptions of individual cases.

This example raises a central issue in model evaluation. Suppose that on a 100-point scale, we can predict decision outcomes on average to within 20 points. Is that good or bad forecasting? Is a 20 point average error a brilliant success, given the complications of political life, or is it a routine accomplishment that should surprise no one? Implicitly, the question asks whether an average error of 20 points is better or worse than a naive or atheoretical forecast. Thus formal political models need a baseline against which to compare their predictive accuracy.

Experience with baselines is commonplace in other disciplines where prediction is important. Meteorology is a particularly instructive example. During most of the twentieth century, meteorologists possessed impressive mathematical theories of how weather patterns occurred. Nevertheless, with the data they had available, their forecasts were not very good. Forecasts were often compared against a simple “persistence” baseline: “The weather tomorrow will be like the weather today.” Another common baseline was “climatological”: “The weather tomorrow will be like the usual weather at this time of year.” Only in recent decades, with the advent of weather satellites and other improvements in data gathering, plus more powerful computing, meteorologists have been able to out–perform baseline models, at least in short–term forecasts covering the next few days (Murphy and Winkler, 1984).

Economists have faced similar forecasting difficulties, and they, too, have
paid attention to proper baselines. Beginning in the 1960’s, ever more com-
plex models of national macro–economies were constructed, sometimes in-
volving many dozens of equations. These models were estimated with his-
torical data, then used to forecast. Forecasting accuracy was often rather
poor. Eventually, good baseline models were constructed using atheoret-
ical time series statistical techniques. The baseline models were vector
autoregressions—essentially an econometric version of the persistence model.
(“The economy next quarter will be like the economy has been lately.”) The
baseline models often outperformed their more sophisticated theoreti-
cal counterparts (Granger and Newbold, 1986: 287-94).

The science of evaluating forecasts has learned much from experience
with meteorological and macroeconomic models. In some instances, their
baselines can be borrowed without modification. For example, in budget
studies, last year’s budget plus inflation is an excellent baseline for assessing
budget forecasting models. In the quieter areas of politics, where essentially
the same decisions are repeated annually on a more or less routine basis,
finding a good baseline is relatively easy.

Unfortunately for most studies of political decisions, however, borrowing
a baseline from meteorology or economics is impossible. Neither the clima-
tological nor persistence analogs works well in politics generally, since public
decisions come along steadily in political life, but they virtually never repeat
themselves in the same form or with the same meaning. When a new public
problem enters the political arena, as they frequently do in the dataset used
in this book, a prediction that “The EU will do the same thing that it did
last year” or that “The EU will do what it always does” is meaningless. For
political baselines, analysts must turn elsewhere.

**Downsian and Related Baselines**

Baselines for political predictions must take into account the specific features of the issue at hand. The most prominent feature of any political issue is the set of policy preferences held by the key political actors. The task for any political model is to use those preferences, along with the intensities that the actors bring to the issue, to predict what will happen within a given institutional structure. A good baseline model would generate predictions of that kind, using only the simplest, most transparent logic.

The most popular choice for a baseline prediction is the Downsian model (Hotelling, 1927; popularized by Downs, 1957). Choices are assumed to be made by round–robin majority rule, whereby each alternative is matched against every other alternative, and the one that defeats all the others is the winner. Alternatives are assumed to be confined to a single dimension. That is, each of the alternatives must be positioned from left to right on an ideological scale, and all voters must agree on their locations. Voters themselves must occupy a position on the same scale, and when given a vote, they must select the alternative closest to themselves on the dimension. Then, as is well known, round–robin majority rule voting always produces a single winner, and that winner is the preference of the median voter.¹ This model is so simple and so well known that in many empirical applications to decision making under majority rule, the Downsian predictions are simply assumed to be correct. Often, no other possibilities are even considered.
The Downsian model also applies to decision making under weighted majority voting. If some voters have more than one vote, then the preferences of the weighted median voter are determinative under majority rule. The logic is virtually identical to that of the unweighted case. The Downsian analysis applies to these and other situations in which the decision is made by all actors jointly and simultaneously, using some version of majority rule.

More complex decision making structures, in which a qualified majority or unanimity are required (such as in the Council of Ministers), in which some actors may have agenda control (like the EU Commission), different subgroups of actors may vote separately (as when a parliament has an upper and lower house), or in which bargaining may be used to settle some issues (as when conference committees must reconcile dissimilar decisions by two different voting bodies)—these do not have the simple form that the Downsian logic requires. In such cases, the median voter’s preference is often not the equilibrium solution, and game-theoretic solution concepts such as subgame perfection are used to generate the predicted outcome of the extensive form game. Then the median or weighted median has only an informal justification, as an approximation, perhaps rather crude and inaccurate, to the outcome of the voting game (Krehbiel, 1998: 232).

In a similar way, the mean of all the actors’ positions has a rough-and-ready interpretation as an approximation to complex decision making procedures. Caplin and Nalebuff (1988, 1991) showed that if a voting rule is 64% majority rule (or more), and if certain additional requirements on the distribution of voter opinions are met, then the mean cannot be defeated when paired against any other alternative. In general, many alternatives will
be undefeated; the Caplin–Nalebuff result is that the mean will be among them.

Since the Council of Ministers has always used a weighted voting rule requiring more than two-thirds majorities, the 64% result has some appeal for forecasting EU decisions, at least as one forecast among others. The difficulty, though, is that the EU does not use a simple supra-majority rule to make decisions because the Council does not act alone. In the presence of agenda control, multiple decision making bodies acting sequentially, and all the other complexities of EU decision making, the 64% rule has no direct logical application to EU decisions. It may function adequately as an approximation for some purposes, but its theoretical relevance is questionable. Like the median, the mean has only an informal justification.

In summary, for the forecasting of governmental decisions, the median and the mean behave like the atheoretical baselines used in meteorology and macroeconomics. That is, they do not embody the theory appropriate to the outcomes being predicted. Moreover, a brief look at the outcomes in the data set used for this book will demonstrate that both the median and the mean are far from perfect predictors. In that respect, too, they parallel the meteorological and econometric baselines. However, the median and mean do incorporate a substantive intuition that is known to apply to other, simpler cases of political decision making. In that sense, they are logical starting points for our analysis of predictive success. A fair test of a model is that it do better than very simple baselines which are known not to be theoretically correct.

For this chapter, a variety of other baselines were also tried. For example,
three weighted medians (with capabilities, saliences, and the product of capabilities and saliences as weights, respectively) were examined, but they proved very slightly less successful on average than the simple unweighted median, and so they were dropped. Similarly, the actor mid–range (the point equidistant from the two most extreme actor positions) proved to be about as accurate as the median and not as good as the mean. Since it has less theoretical justification than either of them, it, too, was dropped. For the purposes of this book, therefore, the baseline models we evaluate are the unweighted median and the mean, each of them computed across all actors, with the Commission and the European Parliament each treated as a single actor. With these two baselines used as the standard of comparison, the models of this book may be evaluated for their forecasting power.

Measures of Model Success

The topic of model evaluation and comparison has generated a vast literature in statistics. (A classic discussion in econometrics is Gaver and Geisel, 1974.) Governments, too, pay attention to political decision making models and conduct their own evaluations. The American Central Intelligence Agency once became quite enthusiastic about Bueno de Mesquita’s Challenge Model, which is set out in Chapter 5 (Feder, 1987).

Yet all such assessments are riven by disputes. In its simplest form, the debate centers around this question: Given a series of models, and given a data set to which all the models apply, which model fits best? Deep conceptual puzzles surround the topic, beginning with questions about the
meaning of the word “best.” Best for what? No general agreement has emerged.

Within particular statistical traditions, the issue of model comparison can sometimes be formulated clearly and sharply. Among Bayesians, for example, the best model is simply the one that has the largest subjective probability, given the data. (A recent brief review is Robert 2001, chap. 7.) Computational challenges and intellectual difficulties arise in executing this program, particularly in defining the initial state of ignorance from which the data are meant to elevate us. Nevertheless, the goal is clear, the fundamental conceptual apparatus is agreed on, and practical conclusions are often unaffected by the details of implementation. In many respects, Bayesian analysis would be the most attractive way to proceed, and it would allow us to make statements such as, “The mean has a 15% chance of being the best model for this dataset, while the median has only a 2% probability.” Thus the predictive prowess of models could be ranked, and the chance of ranking errors would be made explicit. See Raftery (1995) for a simple Bayesian model evaluation procedure that has received attention in the social sciences, and Bartels (1997) for a substantively sophisticated Bayesian evaluation of competing statistical models arising in political science.

For our purposes, however, the difficulty with all this literature, Bayesian and non–Bayesian, is that the models in this book are deterministic. That is, they lack the probability forecasts that are the starting point for statistical model comparisons. A deterministic model may forecast an outcome of 60, while the actual result is 75. Is this “close”? The answer to that question depends on how likely the outcome 75 is under the model, but
no such answer is available. The prediction is simply 60, *tout court*. In that sense, deterministic models in the social sciences are always scientifically incomplete. None of the standard model evaluation procedures can get underway.

Now of course, it is always possible to impose a stochastic structure on a deterministic model. Typically, an additive error term is appended, so that the model becomes incorporated into a statistical specification:

\[
\text{Outcome} = \text{Model Forecast} + \text{Error} \quad (1)
\]

Then by adding some assumptions about the form of the errors, a likelihood function can be generated. Normal (Gaussian) errors are often assumed.

The challenges of this approach are twofold. First, the error structure is not part of the original model. It is tacked on by another analyst. The new statistical assumptions must be consistent with the original structure if the model is not to be penalized for inappropriate additions made by someone else.

Equally important, the choice of the error structure must respect the form of the data. For example, in the data set used in this book, many of the decisions were dichotomous in character (yes/no). If we code yes = 100 and no = 0, then what does a forecast of 70 mean? Should we round it to 100, so that the forecast is “yes” with probability 100%? Or does 70 mean that “yes” will occur with probability 70%? Or is the prediction 100 with some allowance for measurement error in the data, so that the actual forecast is, perhaps, 85% chance of “yes”? (See Chapter 4, Appendix 3.)
Each of these choices could be defended, and statistical analysis could be undertaken to choose among them. For now, however, we have resisted the temptation to impose an elaborate and debatable statistical structure on the models. In consequence, we do not give a full Bayesian comparative analysis of the models. Instead, we adopt a more flexible quantitative approach, evaluating the models on several different statistical criteria.

In the absence of a statistical specification for each forecasting model, the problem of choosing a criterion for model evaluation is analogous to the problem of choosing a voting rule. Think of the forecasting models as the set of candidates standing for election. Treat the individual EU issues as the voters, with each issue counting equally. Each outcome corresponds to that issue’s most–preferred point. Then using the issues to select models near the outcomes is exactly parallel to having individual voters select the candidates near their most–preferred points. In each case, one needs a voting procedure.

Arrow’s Theorem (Arrow 1963) tells us that no voting rule will satisfy all ethical criteria. In parallel, no evaluation criterion will meet all our demands. Nevertheless, the analogy of voting rules suggests that we should look for evaluation rules that resemble sensible voting rules. For example, we could adopt a utilitarian voting rule, in which the sum of individual losses from the candidate choice is minimized. Or we could use the “method of majority decision” (round-robin majority rule), in which each candidate is matched against every other candidate one at a time, and the candidate who defeats all other candidates by simple majority rule is declared best. We could also try less familiar rules such as approval voting (Brams and
Fishburn 1983), in which each voter selects all candidates acceptable to him or her, and the candidate is declared best who has the most such “approval votes.”

The following sections translate these and other voting rules to the context of model evaluation. This latitudinarian style lets us compare models without imposing arbitrary statistical assumptions, and it brings insights of its own. Best of all, it turns out that the ranking of models is little affected by choice of voting rule.\(^4\)

**The Issue of Free Parameters**

One additional area of concern needs to be addressed before the model evaluations are presented. The models in this book all operate without conventional “free parameters,” unknown constants that can be adjusted to improve their fit. On closer inspection, however, most of the models have made use of something equivalent. Thus for example, we have measured power three different ways—two versions of the Shapley–Shubik index and one set of expert opinions. When the second version of the Shapley–Shubik values worked best, that became the definition of “power.” Models that require power measures, such as the Compromise Model (and models like the Exchange Model, the Coalition Model, and the Issue Line Model, which make use of the Compromise Model in their calculations) thus have in effect a free parameter available to improve their fit to the outcomes.

Some models have additional “degrees of freedom” as well. As discussed in Chapter 5, the Exchange Model can be run for one or more rounds. The
number of rounds is a free parameter, and that parameter was chosen in combination with the power measure to maximize the fit. This gives the Exchange Model a total of two free parameters.

The procedural models discussed in Chapters 3 and 8 thus are at a disadvantage, since they make no use of power measures. However, the disadvantages are not all one-sided. Most of the procedural models make no forecast at all on two dozen or more of the 162 issues, so that their errors are based on a different sample from that used by the other models. In effect, they are free to choose the issues they will forecast—a substantial benefit. For example, if the Compromise Model were allowed to drop the 30 cases in which no bargaining agreement was reached, its average prediction error would drop from 22.9 to 20.1, making it an easy and substantial victor over all the other models in this book—so long as they were forced to predict the full 162 cases (see Chapter 4). In effect, picking one’s cases for prediction is a free parameter, too. In this respect, the procedural models are advantaged.

All these considerations make the model error comparisons inexact. It is not clear how severely additional free parameters should be penalized, nor whether models with missing predictions should be downgraded, and if so, by how much. Moreover, should we credit most of the procedural models for their parsimony in ignoring the intensities of the actors, or discount them because they do not have to cope with the measurement errors in the saliency measures? Rather than attempting to adjudicate all these deep problems of model evaluation, we simply present the average forecast errors for those issues that a given model is able to forecast. And in the interests of scientific honesty, each chapter has presented the results of alternate choices.
For example, Chapter 4 sets out how well the Compromise Model performs under different measures of power, and Chapter 5 does the same for the Exchange Model, assessing the impact of both the power measure and the number of rounds the model was run. Thus the reader is free to compare not only the findings of this chapter, which gives the fit of the best models, but also the performance of alternate versions of each model. In general, these comparisons show that the real differences in forecasting power occur across models, not within them. That is, different versions of the same model have very similar predictive power. Hence the model evaluations of this chapter are sensible.

Mean Absolute Error per Issue: The MAE

We now take up the model evaluations themselves. The first measure is the simplest, easiest to understand, and perhaps the best overall—the mean absolute error (MAE). The reader has already encountered it in previous chapters. The MAE is just the average size of the forecasting mistake. Choosing models for smallest average error corresponds to a utilitarian voting rule. That is, maximizing total satisfaction of the citizens is the same as maximizing average satisfaction, and that in turn is the same as minimizing average loss. Utilitarian ideas have a long history in ethics and economics, and their analogs for model evaluation are a natural starting point. Minimizing average loss for citizens corresponds to minimizing the average forecasting error for models.

Thus for outcomes $y_i$ ($i = 1, \ldots, n$), and a model $M$ with corresponding
forecasts $\hat{y}_i$, we define the MAE as:

$$\text{MAE}_M = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$  \hspace{1cm} (2)

This measure has the advantage of answering a simple question: How far wrong is the model, on average? Of course, to make the measure sensible, all issues must be scaled to the same range. In this book, that range is taken to be 0 to 100. Thus an MAE of 30 means, for example, that on average across all issues, the difference between the model prediction and the actual outcome is 30 points. Note that the unit of analysis for the MAE is the issue. All issues are counted equally, regardless of whether they occurred singly in a proposal or as part of a package of several issues included in a single proposal.

In Table 10.1, we show the mean absolute error (MAE) for eleven models from this book, plus the two baseline models (the median and the mean). The table shows only the MAEs for the best model from each of the previous chapters. (The MAE values for less successful models appear in the chapters where they were introduced.) The Challenge Model, which played a prominent role in Bueno de Mesquita and Stokman (1994) and was by some measures the most successful model in that book, is also included in Table 10.1, as is the State–Centric Realist Model, patterned after Mearsheimer (1994-1995), which has received considerable attention in the literature.\textsuperscript{5}

The right–hand column of Table 10.1 conveys an important message: Most of the models considered in this book are not as good at predicting outcomes as either the median or mean baseline models. In fact, the simple
mean is a very strong predictor of outcomes in this data set, and no other model is substantially better. Only the Compromise Model (Chapter 4), the Exchange Model (Chapter 5), the Coalition Model (Chapter 7), the State–Centric Realism Model (Chapter 4), and the Domestic Constraints Model (Chapter 6) are approximately as good as the mean, perhaps because each of them either makes substantial use of means and weighted means in their predictions or else uses the Nash Bargaining Model, which has much the same effect. By contrast, the theoretically powerful procedural models do much less well as a group.

The models break up into four categories of predictive success. The first category includes the Compromise Model, the mean, and the Exchange Model. This triumvirate is the most successful, with the Compromise Model finishing very slightly ahead, but not in a statistically reliable way. All three models “split the difference” among the actors’ preferences in a bargaining or exchange style.

The second group contains models very close in predictive power to the Compromise Model, the mean, and the Exchange Model. In fact, their statistical performance cannot be reliably distinguished from the top group. This set of models includes the Coalition Model, the State–Centric Realism Model, and the Domestic Constraints Model. Each of these models is based on bargaining, exchange, and/or coalition formation. All of them use either the Compromise Model and/or the Nash Bargaining Solution at some stage of their calculations. Like the mean, Compromise, and Exchange Models, none of these takes explicit account of the details of the EU decision rules or legal procedures (apart from the list of winning coalitions specified by the
various treaties establishing the EU).

The third category of predictive success has two members—the median and the Challenge Model due to Bueno de Mesquita. These two models perform distinctly less well than the best six models. (In difference of means tests, these two models lose to all the models in group one at least at the .20 level, and most of the differences are significant at .05.) This second group of models share an emphasis on the median rather than the mean as a predictive tool. They also ignore the details of the legal procedures.

The fourth and final category of models, the least successful, do drastically less well than the models in the first category, with mean errors as much as 50% larger and very significant difference of means tests. They also do somewhat less well than the models in the second group. This fourth group of models includes, for example, all the procedural models based on EU decision making rules (Chapter 3). The Tsebelis model, falls into this highly unsuccessful category, as do other, non-procedural models such as Coleman’s. Like the models in the third category, the models in this third group generally make no use of mean or weighted mean forecasts in their calculations.

In summary, means are better than medians, and both are better than extensive form games.

Table 10.1 also presents MAE values under four different European Union voting procedures—consultation with qualified majority in the Council, consultation under unanimity, codecision with qualified majority, and codecision with unanimity. Broadly speaking, the pattern of model success is similar across these procedures, with the same models doing well each
time. However, as close inspection of the table will show, there are intriguing unexpected successes and failures scattered here and there. For example, the generally unimpressive Coleman model predicts in only half the cases of consultation under unanimity, but when it does so, it is very successful. Whether this and other deviations from average success represent genuine theoretical information, the ease of predicting particular subsets of issues, or just random variation due to small numbers of cases awaits further research.

Table 10.2 presents the same average forecast errors, this time organized by issue area. For agriculture and “other” policy areas, the general pattern holds, with the same models doing well and poorly as in the overall rankings. However, for nearly all the models, the internal market issues seem more difficult to predict than the other two issue categories, and their prediction errors rise proportionately. Two exceptions are the Exchange Model, which does very well on internal market issues, and the State–Centric Realism model, which finishes first in that issue area. Is the internal market the true realm of state power in the EU, while agriculture is not? Doubts about that claim would be buttressed by noting that the internal market contains only 33 issues and that the prediction errors have large variances, meaning that chance plays a substantial role in these comparisons.8 Thus additional detailed investigation, both statistical and historical, will be required to uncover whether the empirical differences in the various models’ success across policy categories are theoretically informative.

Table 10.3 again displays the MAE values, this time categorized by the measurement level of the issue outcome. Dichotomous outcomes have the largest prediction errors, ranked outcomes are next, and interval scales have
the smallest errors, as would be expected from their levels of measurement. However, with some allowance for sampling error, the pattern of errors seems quite similar across the scale types, with the same models doing relatively well and badly on each of them.

There are two important exceptions to the generalization that scale type does not affect the relative performance of the models. The first occurs in comparing the mean versus the median. The mean outperforms the median except on dichotomous issues, where the median’s advantage is substantial. Similarly, the Exchange Model, which is an average performer otherwise, scores a brilliant success on the dichotomous items, where it is easily the best model. By contrast, two other very successful models struggle with the dichotomous items. The Compromise Model performs only modestly well, and the mean finishes well below its usual rank, beaten by eight other models.

Further investigation (not shown) demonstrates that the Exchange Model achieves its advantage on the dichotomous items primarily by doing well on multi–issue proposals. The cross–issue trades emphasized by the Exchange Model probably affect dichotomous items the most, since these items have no intermediate positions. Vote exchanges and trades move dichotomous outcomes a long way on our scales, and the Exchange Model seems to take the best account of that.

One must remember, however, that large forecasting errors are particularly common on dichotomous items: One forecast of 30 when the outcome is 100 adds by itself more than two points to the MAE. Two or three substantial prediction errors occurring by chance among 33 dichotomous
forecasts can easily change a model’s average error by five points or even more, enough to turn one of the best models into one of the worst. Thus as with all the other statistical evaluations, in the absence of additional detailed case studies of the dichotomous issues and careful statistical analysis, interpretive caution is warranted.

**Mean Euclidean Error per Proposal: The MEE**

The second measure of closeness differs from the MAE in two ways. First, to take account of the fact that EU proposals often contain several issues considered jointly, we make the unit of prediction the proposal rather than the issue. Second, we adopt a measure of closeness designed for a multi-issue prediction space, namely the Euclidean distance from the prediction to the outcome. The squared Euclidean distance is divided by the number of issues within the proposal to give the average squared distance per issue, and a square root is taken. This is the notion of root mean squared error, computed in this case within each proposal. Finally, these square roots are averaged across proposals, with all proposals counted equally.

This measure treats all the issues as if they were continuous scales, when in fact many are discrete and even dichotomous. However, treating discrete and ordinal data as continuous has long been statistically defensible as an good approximation to far more complex calculations, and we have adopted that approach here (Abelson and Tukey, 1959). Similarly, embedding the non-continuous issues in Euclidean space and treating the resulting “distances” as continuous is intended to be merely heuristic.
As a voting rule, the MEE makes the proposals the voters, not the issues. It also treats losses as essentially quadratic rather than linear. Thus it corresponds, perhaps, to a “functional representation” voting rule, in which groups, not people, are the voters. The groups, not necessarily all the same size, are then weighted equally in a utilitarian calculation. Thus again this criterion has less appeal than the MAE’s straightforward utilitarian perspective, but it does provide an alternate perspective emphasizing proposals rather than issues as fundamental.

To spell out this definition, suppose that there are \( m \) proposals altogether, each with \( n_j \) \((j = 1, \ldots, m)\) issues included in it. Thus the total number of issues is, as before, \( n = \sum n_j \). Denote the \( i \)th issue in the \( j \)th proposal by \( y_{ij} \) \((i = 1, \ldots, n_j)\). Then we define the mean Euclidean error (MEE) for model \( M \) as follows:

\[
\text{MEE}_M = \left( \frac{1}{m} \sum_{j=1}^{m} \left[ \sum_{i=1}^{n_j} (y_i - \hat{y}_i)^2 / n_j \right] \right)^{.5}
\]

(3)

With this definition, the MEE is a straightforward extension of root mean squared error (RMSE), familiar from engineering and economics applications. Here the RMSE is computed across issues within each proposal, in accordance with the standard definition of Euclidean distance. Then the RMSE is averaged across proposals, with all proposals weighted equally, to give the MEE.

Table 10.4 displays the MEE values for each model. We regard this table primarily as a check on the previous tables, which used absolute values and averaged over issues. Happily, Table 10.4 shows that Euclidean distances
and averages over proposals do not affect the fundamental ranking of the models. To be sure, there are mild differences here from the previous tables due to the altered definitions. But again the mean, the Exchange Model, and the Compromise Model perform best, again with no statistically reliable differences among them. (This time the mean is a bit better than the other two, which are essentially tied.) The Coalition, State–Centric Realism, and Coleman models are the next best performers, close enough that with a larger sample, any of them might prove to be the best performer. The other models are all less successful, as before.

Correlations Simplistic and Sensible

The third definition of closeness used in this chapter is “goodness of fit.” For this purpose, we set out two different correlation measures. The first is the simple Pearson correlation coefficient between each model’s predictions and the outcomes. The Pearson $r$ has the advantage of familiarity and simplicity. Moreover, with every model judged by the same actual outcomes, the correlations generated by different models may be sensibly compared.\footnote{11}

It is important to remember, however, that Pearson correlations do not actually measure “closeness” in the usual sense. They do not capture how well predictions fit the outcomes. Instead, they are measures of how well a linear transformation of the predictions would fit the outcomes. For example, suppose that the actual policy outcomes from three issues are 0, 50, and 100. A first model predicts 0, 50, and 99. A second model predicts 25, 26, and 27. Obviously the first model is vastly better, in the sense of
being nearer the outcomes. However, by the Pearson correlation criterion, the second model is the better model, since it correlates perfectly \((r = 1.0)\) with the outcomes, while the first model does not.

The problem here is that the Pearson correlation allows 25, 26, and 27 to be transformed linearly before they are compared to the outcome. By subtracting 25 from each of them, then multiplying by 50, they turn into 0, 50, and 100, so their fit is perfect. But this result is a considerable distortion of the usual meaning of predictive accuracy. Thus the correlation coefficient can be quite misleading in this context. It also has no meaningful interpretation as an analog to a voting rule. For measuring predictive power, it is the least satisfactory measure among those we use.

Nevertheless, the examination of correlations can be helpful in comparing forecasts. A particular model may be predicting systematically low or high or within too small a range, but if the pattern of its forecasts has a linear relationship to outcomes, that information may be exploitable in subsequent model building.

Table 10.5 displays the Pearson correlations for each model. The Exchange Model is slightly the best by this criterion, primarily due to its advantage on the dichotomous items with their wide range of outcomes. The Compromise Model and the mean are close behind, and the Coalition and State–Centric Realism models once more finish in the fourth and fifth positions, while the other models have lower values. Again, as is well known, correlations are particularly noisy measures, so that the precise order of finish has no special importance.\(^12\) As with all the other tables in this chapter, the only sure conclusion is that the top half–dozen models perform better.
as a group on this criterion than those nearer the bottom. However, these rankings have all the limitations of Pearson correlation coefficients, and we regard them primarily as indirect clues for building better models rather than as true forecasting evaluations.

A more sensible and useful correlation may be defined. To see how, recall that the conventional $R^2$ in regression analysis gives the “percentage of the total variance accounted for” by the model. In that case, “total variance” is the variance in the dependent variable, i.e., the squared error remaining after the mean of the dependent variable alone has been used to forecast outcomes. The usual $R^2$ uses a regression model to compute how much of the remaining error can be accounted for with a linear transformation of the explanatory variables in the model.

To define a new correlation coefficient, we need a starting point similar to the mean of the dependent variable. However, the mean will not work here. The decision outcomes analyzed in this book are recorded on 0-100 scales, each of which has no natural left-to-right orientation. Flipping some of them left to right would not change their meaning, though it would change the mean of the dependent variable. Thus computing their mean makes little sense. Moreover, we do not wish to allow linear transformations of the forecasts. Hence we need to proceed differently in constructing a correlation coefficient.

First, we choose a value in the observations that is invariant under a left-to-right transformation. For that purpose, we select the midpoint 50 on all our issue scales. Obviously, flipping the scale left to right leaves this value unchanged. Thus the starting prediction, the analog of the mean in
the regression case, is just the midpoint of the scale on each issue.

Next, models are evaluated for their ability to improve on the forecast of 50. We will call this forecast the Midrange Model. The average squared error of the outcomes around this “middle prediction” gives us a “total variance to be explained.”\textsuperscript{13} Then finally, we define a “pseudo–$R^2$” for a given model as the model’s percentage reduction in mean squared errors compared to the Midrange Model. No transformation of the model forecasts is allowed: The actual forecasts are used in their original form.

The pseudo–$R^2$ is defined formally as follows. Call the actual outcomes $y_i$ ($i = 1, \ldots, n$), and let the predicted values from the Midrange Model be denoted by $\hat{y}_{MM}$. Then the sum of squared errors for the Midrange Model is $SS_{MM} = \sum_{i=1}^{n} (y_i - \hat{y}_{MM})^2$. The sum of squared errors for any other model $M$ is denoted by $SS_M$ and is defined in parallel fashion. Then the pseudo–$R^2$ for model $M$ is defined as:

$$R^2_{\text{pseudo}} = 1 - \frac{SS_M}{SS_{MM}}$$

Thus starting from the Midrange Model, the pseudo–$R^2$ gives the additional fractional reduction in explained sum of squares achieved by a given model. Note that, unlike the usual $R^2$, the pseudo–$R^2$ can take on negative values if a model performs worse than the Midrange Model.\textsuperscript{14}

In addition to its advantages over the Pearson correlation coefficient, the advantage of the pseudo–$R^2$ compared to simple MAE or MEE measures is that it takes account of the level of measurement. For example, dichotomous measures convey less statistical information and thus are more difficult to
predict than interval–level measures, so that larger errors on the dichotomous items are likely for all models. Expressing the squared errors for each model as a percent of the squared errors for the Midrange Model is one way to adjust for the differences in scale type.

The pseudo–$R^2$ values of the best models from previous chapters are given in Table 10.6. As a glance at the table will show, only a few models did as well or better than the Midrange Model. The Compromise Mode, the mean, and the Exchange Model have pseudo–$R^2$ values of just under .2, with the Compromise Model performing slightly better than the other two. The Coalition and State–Centric Realism Models are fourth and fifth best, with pseudo–correlation values of about .1. All the other models in this study perform less well than the Midrange Model, and therefore they have negative pseudo–$R^2$ values. That is, they perform less well than a model that simply predicts the value 50 all the time.\textsuperscript{15}

Again, not too much should be made of the exact order of finish in Table 10.6. The point is rather that we arrive at the same relative ranking of models by this criterion as by previous measures: the top five models do better than those further down the list. Nor is this finding a surprise. The main point of the pseudo–$R^2$ measure was to adjust for scale types. But Table 10.3 showed that the ranking of models is very little affected by the scale of measurement. Models that did well on the interval–scale issues generally did well on the ranked and dichotomous issues also, and those that did poorly on one type of measurement did poorly on the others as well. Hence the adjustment for scale type has only small marginal effects.
Pairwise Comparisons of Models

The next measure of model performance is comparative. We match each model pairwise against every other model. For each pair, we determine the proportion of issues on which the first model is closer to the actual outcome than the second model. That is, we compute the proportion of issues on which the first model is a better forecast. This fraction may then be evaluated statistically by the sign test, and statistically significant differences between models can be distinguished from those that are more likely to be due to chance.

Regarded as a voting rule, this criterion for model evaluation corresponds to the method of majority decision, sometimes called round–robin majority rule. Under that procedure, as noted above, all candidates are matched against each other pairwise. The candidate, if any, who defeats all the others in one–on–one contests by majority rule is declared elected. Now for model evaluation, it is the issues that have “votes.” They determine whether there is a model that in pairwise competition defeats all the other models. A winning model must enter every one–on–one competition against all the other models and defeat them in the sense of being closer to the actual outcomes on a majority of issues. Using this ancient voting procedure has an obvious appeal as a criterion for model evaluation.

The central advantage of this measure is that, like majority rule and unlike utilitarian rules, it makes use only of ordinal information on each issue. It asks: Which of these two models was closer in predicting this decision? Essentially, the measure just counts which model won each issue,
and then computes a winning percentage. Thus this measure is relatively robust and quite conservative in its assumptions. Its weakness, of course, is that it counts a close win the same as a big victory, and thus loses some information relative to the MAE and MEE. But used in conjunction with other measures, it brings its own strength, which is very limited dependence on the type of scale used for each issue.

Tables 10.7 computes the pairwise winning percentages among the models, where a “win” is a prediction closer to the outcome than the other model being compared. Statistical significance is computed for each pair using the sign test.

Again the ranking of the models is very similar to that of the other tables. The Compromise Model is the only one with a winning percentage above 50% against every other model. However, it just barely defeats the mean and Coalition Models, and has success rates of approximately 55-60% against the others. Thus the Compromise Model wins, but not in a landslide.

The other models fall into a near–rank order in the comparisons. Thus the mean loses only to the Compromise Model and defeats everyone else, the Coalition Model loses only to the Compromise Model and the mean, State–Centric Realism and the Exchange Model are next best, down to the Issue Line Model, which loses to everyone.17 Thus the comparisons form almost a (stochastically) transitive scale, with only a few close reversals in the middle. And again the implied ranking is essentially the same as in all the previous tables. Obviously, however, each successful model’s margin of victories is small and subject to substantial randomness. The significance tests again show that the same three groups of models—best, middle, and
Hit Rates

We now take up the “hit rate,” that is, the percentage of time that a model forecast, rounded to the nearest possible answer, was correct. For example, on dichotomous issues (scored either 0 or 100), a hit was scored if the model prediction fell on the correct side of 50. A similar procedure was used for the ranked scales. On interval scales, a hit was scored if a model came within ten points of the actual outcome.

This criterion for model evaluation corresponds to an approval voting rule, in which voters are allowed to select as many candidates as they approve of, giving one vote to each. The winner is then the candidate with the most votes. In the same way, in counting hits, each issue “approves” a model if its forecast is near the outcome. Issues can approve as many models as they like, giving a hit to each one. The model with the most hits is the best model under this criterion.

Obviously this measure, like the corresponding voting rule, is a rather special criterion lacking the long-standing appeal and history of utilitarian or majority rule procedures. Nevertheless, we include it here because it may detect a model that behaves like the proverbial little girl: When she was good, she was very, very good, but when she was bad, she was horrid. That is, the hit-rate measure favors models that are sometimes very, very good at prediction, even if they are bad, or even horrid, at other times.
Despite their errors, models with occasional dramatic successes can be quite informative for the next round of model building.

Table 10.8 gives the “hit rates” for each of the models. Approximately the same overall ordering of the models holds once again. Now the mean finishes first, with the Exchange Model, the Compromise Model, the State–Centric Realism Model, and the Coalition Model all very close behind. Of course, none of the differences among these five models are statistically significant, or even close to significance. Thus once again no special meaning attaches to the detailed order of finish.

As it did on the pseudo–correlation measure, the Domestic Constraints Model falls out of the first group and into the second category, since it performs only a little better than the Expected Utility Model and the median. Interestingly, too, by this measure the procedural models rise up to the second tier of models. In fact, they all out–perform the median. Is this a hint that procedural models sometimes get the outcome very nearly exactly right, but other times miss badly? Does that happen because strict legal procedures control some decisions, while bargaining dominates others? If so, there is more than one statistical “regime” in EU decision making, and procedural models would be expected to do fairly well on “hits” and not very well on mean absolute forecast errors, just as they do in the decisions analyzed in this book. Detailed statistical and case study analysis would be needed to answer these questions, but the effort seems worthwhile.
Which Models are Best?

The five measures discussed thus far—the MAE, the MEE, the pseudo-correlation, the pairwise win percentages, and the hit rates—comprise the model evaluations undertaken in this chapter. We make no overarching claims for any of them. Each of them would make excellent sense statistically if certain distributional assumptions held for all the issues (for example, normal distributions for the MEE), but no such assumptions can apply to the variety of issues and levels of measurement present in the data. Instead, we adopt these five measures because each has substantial intuitive content. Their meanings are clear and familiar, as are their limitations. Thus our approach to model evaluation is heuristic rather than rigorous. Happily, it turned out that the conclusions of our analysis did not depend on the choice of evaluation measure that we employed.

In summary, all our model comparisons come to essentially the same result. The Compromise Model, the mean, and the Exchange Model perform somewhat better in our comparisons than their nearest competitors, with typical absolute errors of about 23 points on a 100 point scale. They also have predictions closer to actual outcomes in head-to-head competition with other models, along with higher “hit rates.” These three models, closely related in their theoretical workings, monopolized the top three positions on nearly every overall test we performed.

Forced to choose a first-place model, one might say that the Compromise Model wins narrowly. It comes in first on three of the five relevant measures (the MAE, the pseudo-correlation, and the pairwise comparisons). Two of
these victories occur on important criteria analogous to the utilitarian and majority–rule voting procedures (MAE and pairwise comparisons).\(^\text{18}\)

The mean finishes second. It captures two top prizes analogous to weighted utilitarian and approval voting analogs (the MEE and the hit rate), and is second on the other three measures. The Exchange Model is third. In the model evaluations, it finishes second twice, third twice, and fourth once, yet always close to the top, and it wins the (less relevant) Pearson correlation. However, these differences are all small, statistically unreliable, and not independent of each other, meaning that the outcome among the three top models is best described as a virtual tie.

The Coalition and State–Centric Realism models also predict well, and the Domestic Constraints Model frequently does so—a difference in predictive power from the top three models that is occasionally statistically significant but not uniformly so. The median and Challenge models typically finish next best in our tests. All the other models generate more dramatic errors, and they also fail to predict on many of the issues. They are statistically clearly inferior to the best models.

As with the top three models, the evaluations of lower–ranked models do not change if Euclidean errors on multiple issues are used in place of absolute errors on single issues, nor are they altered if we average by EU proposal rather than by issues. Correlational measures, bilateral competitive tests, and point prediction “hit rates” generated only small differences in model evaluations, none of them statistically reliable. By and large, too, the rankings are the same in different issue arenas and for different levels of measurement.
Thus we have some confidence that the overall rankings of the models reflect genuine differences in predictive power. In particular, the top five or six models virtually always rank best under any plausible measure of predictive success, with the Compromise Model, the mean, and the Exchange Model almost always taking the top three positions. That is the central finding of the model comparisons.

The remaining topic, then, is the theoretical inference to be drawn from these comparisons. We also take up the question of how our findings should influence future theorizing.

Conclusions

Formal mathematical models in political science have become less abstract and more detailed in their description of institutions. Increasingly, they can be used to predict both quantitative outcomes in laboratory experiments and qualitative features of political life (Morton, 1999). Some formal models, such as those discussed in this book, go further and generate precise predictions of actual political events, such as decisions by policymakers.

The comparative statistical study of such models stems from Bueno de Mesquita and Stokman (1994). The present work has followed in their footsteps and attempted to advance the subject by exploring a variety of additional models and evaluating them for predictive accuracy with a much larger data set. What do these model evaluations teach us about our current theories? Five central conclusions seem to arise.

First, social scientists are very far from predicting political decisions
accurately. Even our best models have average errors exceeding 20 points on a 100 point scale. To see how modest that level of predictive success is, suppose that outcomes occurred randomly and uniformly across the 100 points of the scale. Suppose further that we forecast a random point on the scale, also distributed uniformly and independently of the actual outcome. Such forecasts are obviously completely useless, and they correlate exactly zero with the outcomes. Thus they form a useful lower bound for comparison of prediction errors.

Now a little calculation shows that the mean forecast error (MAE) for this useless random model would be 33.3. Alas, most of the theoretically sophisticated procedural models discussed in this book, based on extensive form games, are performing only slightly better than that, with MAEs of 30 or more. Even our best models cut that error rate only by a third, to about 23. Clearly, on any absolute scale of predictive accuracy, we have far to go. Our models differ from reality far more than they differ from each other. For example, the forecasts from the Compromise Model and the Exchange Model correlate with each other at .87, but each correlates less than .5 with the actual outcomes.

Nor is that finding solely a judgement on formal models. Neither case studies nor statistical modelling have pointed the way to better predictions. In EU studies as in political science as a whole, we are far from having the conceptual tools of any methodological type that we need to forecast political decision making well.

Second, as a group, the procedural models based entirely on the legal rules of EU decision making do not perform well on average. However EU
decision making is carried out, it does not seem to be well described solely
by the formal rules. Informal norms and procedures appear to play a more
central role. Seen in this light, the prominent debate between Tsebelis
(1994) and Crombez (1996) over the proper extensive form for modelling
the EU emerges as a rather minor concern. The forecasting weaknesses of
extensive form games stem from deeper problems.

Third, the state-centric realism proposed by Mearsheimer (1994-1995)
also proves inadequate. This approach, in which national interests deter-
mine the behaviour of international institutions, does not fare badly in our
statistical tests, but it is consistently outperformed by similar models that
incorporate the postulates of liberalism. In particular, assuming that inter-
national organs such as the European Parliament and the Commission play
an independent role beyond state interests leads to better forecasts.

Fourth, the reference (or reversion) point does not seem to play the cen-
tral role in decision making that extensive form games imply that it should.
Consider, for example, those issues on which the decision outcome was far
from the reference point. (Often the reference point is the status quo, so
that outcomes far from the reference point result in a large change from
current policy.) These are issues in which the reference point was weak.
It was overcome in some way and did not constrain the decisionmakers to
locate nearby. In Table 10.9, decisions of that kind are treated as predictors
of model errors. More precisely, the size of the outcome change from the re-
ference point (denoted “Change”) is correlated with each model’s predictive
errors.

Table 10.9 shows that the procedural models, many of which make cru-
cial use of the reference point in predicting errors, tend to make larger errors when the reference point is weak in influence. That is, they put too much weight on it. By contrast, the other models ignore the reference point, they do better overall, and their prediction errors do not generally grow worse when the decision outcome represents a large change from the reference outcome. More detailed investigation will be necessary to fully validate this claim about the over-emphasis on reversion points in the procedural modelling tradition, but Table 10.9 provides some evidence that coalitions and bargaining somehow downweight the leverage that reversion points logically seem to have in our most powerful theoretical models.

Fifth and perhaps most important, the models that are most successful all compute some sort of mean among actor preferences, weighted or unweighted. In these models, EU decision making is treated as if extreme positions are accommodated rather than being ignored, even when ignoring them would still allow the other actors to get their way. Moreover, there is some evidence from the Compromise, Exchange, and Coalition Models that power and salience also matter. That is, powerful and intense actors are conciliated, even when they might legally be ignored. By bargaining and/or exchange, actors may give up certain goals they care less about in exchange for other goals they value more.

The conclusion that bargaining and compromise are central to EU decision making will come as no surprise to political practitioners and participant observers. The case study literature has repeatedly emphasized the role of compromise and the striving for unanimity in EU decision making. States are disinclined to follow the letter of their legal rights if doing so makes an
enemy. Bargaining matters more than the official decision making rules.

In the data analyzed in this book, too, the outcomes often show clear tradeoffs across issues, with states that care less about particular topics deferring to more intense states, while the less intense states in turn get their way on other issues of greater concern to them. But such tradeoffs do not occur in most procedural models, which typically analyze issues one decision at a time. By its nature, then, a single-issue analysis of the sort used by many conventional procedural models will struggle to capture the cross-issue trades that political actors make. Put another way, procedural models usually ignore intensity differences across different decisions simply because they look myopically at each decision on its own. The result is that they may be beaten in predictive accuracy, as they are in this book, by less theoretically sophisticated models that take those intensities into account. The inference would seem to be that good forecasting requires that bargaining and intensity be part of the model.

It is important not to misread the evaluations in this chapter. It would be quite wrong to infer that procedural models should be discarded in favor of very simple models like the mean and the Compromise Model, or any of the other top-performing models of this book such as the Exchange and Coalition Models. None of these successful models takes real account of the ways in which institutional forms and legal structures create actors’ strategy spaces. Only procedural models do that, and in so doing, they have taught us too much to throw them away.

The lesson instead is that procedural models need theoretical extension. They need to take into account, not just the formal rules, but also the
informal processes that make up so much of what politicians do. Creating theoretical structures adequate to that challenge will not be easy. But combining bargaining and vote trading theory on the one hand with the procedural framework on the other, and doing so in the theoretically coherent and defensible manner that is the hallmark of modern game theory—that, surely, is the future of forecasting political decisions.
Notes

1I am grateful to the Center for the Study of Democratic Processes, Princeton University, and the Department of Political Science, University of Michigan, who provided released time from teaching for this research. A research fund at the University of Michigan donated by Norma Shapiro supported the overseas trips so important to international collaboration, and I thank her for her help. I also thank the authors of the other chapters in this book plus Simon Hug and John Jackson for many constructive comments and suggestions. Marcel van Assen, Larry Bartels, and the graduate students in Politics 583 at Princeton in Spring Term, 2004, also gave helpful additions and criticisms, as did seminar participants at the University of Michigan, National Chengchi University of Taiwan, and Rutgers University. Robert Thomson and Jacob Dijkstra skillfully carried out the calculations and produced the tables. Remaining errors are my own.

2This statement assumes an odd number of voters, so that there is only one median voter, but with obvious modifications, the statement holds for an even number of voters as well.

3Weighted means are also represented in the book, though not as baseline models: The Base Model and the Compromise Model forecasts, both discussed in Chapter 4, are just weighted means, with capabilities and the product of capabilities and saliences as their respective weights.

4In addition, as mentioned in Chapter 2, the data seem free of systematic distortions that might distort the model rankings. For example, when we examine just those issues whose expert respondents were affiliated with the permanent representatives of the member states, the accuracy of the models does not differ significantly from those issues whose experts had a different affiliation. Thus national biases in the data themselves are not worrisome, and we can proceed to the evaluations.

5For purposes of comparability, the version of the Challenge Model assessed here is the same as that used in Bueno de Mesquita and Stokman (1994). Modifications to the Challenge Model that may have been made since the first book was published are not included. This model and the Exchange Model are not in the academic public domain; they are proprietary.
Difference of means tests (paired observations) for these data show that two models must have MAE values that differ by about 3.5 or 4 to be statistically significant at the conventional .05 level. (The precise difference needed varies slightly from one model pair to another.)

The worst models in this third group are statistically distinguishable at the .05 level from all the models in group two, and the Tsebelis model is nearly so. The other procedural models (from Chapters 3 and 8) differ from group two only at approximately the .25 level, which of course is not a reliable difference. However, if we add the consideration that all the procedural models fail to forecast in nearly one third of the cases, it seems fair to place them below the first two groups in overall predictive power.

In addition to the usual caveats, it is worth remembering that a t-test comparing a model’s best performance to its average performance exaggerates the statistical significance of the difference. By picking the best performance to compare, the researcher has capitalized on chance in a way that violates the assumptions underlying the test. Put another way, we would expect to see some substantial statistical differences in model performance across issue types just by chance, even if all were due to random noise.

In particular, since the Exchange Model and the Compromise Model make identical forecasts on single–issue items, the Exchange Model’s advantage over the Compromise Model on dichotomous items is entirely due to multi-issue proposals.

Note that without the division by the number of issues in each proposal, those proposals with large numbers of issues would dominate the average, violating the goal of this measure, which is to average over proposals rather than over issues.

Comparing correlations across different samples is much harder, if not impossible, as good statistical texts remind readers.

This is particularly true in the present case, with non–normally distributed data and thus an unknown sampling distribution for the correlations.

Of course, this quantity is not a variance, hence the quotation marks.

The pseudo–$R^2$ also has an interpretation as an (unweighted) utilitarian rule that uses quadratic losses as its utility measure.

Other pseudo–$R^2$ definitions might be proposed. For example, instead of averaging
over issues, we might wish to average over proposals, using the MEE values from Table 4. The new pseudo-\( R^2 \) would then be defined as in Equation (3), with each model’s MEE replacing the SS values. A glance at Table 4 will demonstrate that the same models do well and poorly under this definition as in the other comparisons of this chapter.

16 This is not quite true. If the two model forecasts fall on opposite sides of the outcome, the spacing of ordinal items on the underlying scale matters in determining which forecast is “closer.” However, it is certainly true that the sign test is much less dependent on interval-level assumptions than our other measures of model accuracy.

17 The Exchange Model’s highly focused success on the dichotomous items, combined with its average performance on other scales, means that it will do less well on these paired comparisons, which average across all issues, than it does on hit rates, which reward great success on a minority of items.

18 In the original plan of this book, the Compromise Model was to serve as a baseline to demonstrate the additional predictive power of other models. No theoretical development of the model was planned because no need for it was foreseen. When the Compromise Model unexpectedly performed as well as it did, however, the surprised author of this chapter turned his hand to developing the argument for the Compromise Model set out in Chapter 4.

19 The Domestic Constraints Model also employs the reference point in its calculations, but only as part of a Nash bargaining solution. That is, it uses the reference point to establish bargaining weights, but it does not employ an extensive form game that treats reference points as a crucial factor in actors’ thinking about their decisions. Perhaps for that reason, the errors of the Domestic Constraints Model are negligibly correlated with the variable Change.

References


395