Routledge
Taylor & Francis Group

Critical Review

ISSN: 0891-3811 (Print) 1933-8007 (Online) Journal homepage: http://www.tandfonline.com/loi/rcri20

JUDGING JUDGMENT

Bruce Bueno de Mesquita

To cite this article: Bruce Bueno de Mesquita (2010) JUDGING JUDGMENT, Critical Review, 22:4, 355-388, DOI: 10.1080/08913811.2010.541686

To link to this article: http://dx.doi.org/10.1080/08913811.2010.541686

Published online: 26 Feb 2011.

Bruce Bueno de Mesquita

JUDGING JUDGMENT

ABSTRACT: *Philip E. Tetlock and I agree that forecasting tools are best evaluated in peer-reviewed settings and in comparison not only to expert judgments, but also to alternative modeling strategies. Applying his suggested standards of assessment, however, certain forecasting models not only outperform expert judgments, but also have gone head-to-head with alternative models and outperformed them. This track record demonstrates the capability to make significant, reliable predictions of difficult, complex events. The record has unfolded, contrary to Tetlock's contention, not only in government and business applications, but also in numerous peer-reviewed publications containing hundreds of real-time forecasts. Moreover, reliable prediction is achieved while avoiding significant false-positive or false-negative rates.*

Expert political judgment, as Philip Tetlock (2005), Richard Posner (2002), and others have shown, is a poor source of reliable predictions. It seems surprising that anyone would have thought otherwise. What, after all, might we reasonably expect is the "expertise" of area experts, subject experts, or problem experts: facts or judgment?

We might wish for such experts to have excellent judgment, discernment and wisdom—and some surely do—but these abilities are not needed to be an expert and certainly are not part of experts' training. Expertise consists of an unusual accumulation of knowledge and facts regarding a subject. It is not about having a reliable means of translating

Bruce Bueno de Mesquita, bbd2@nyu.edu, Julius Silver Professor of Politics, New York University, New York, NY 10012, is the author, *inter alia,* of, *The Predictioneer's Game* (Random House, 2009).

Critical Review 22(4): 355–388 ISSN 0891-3811 print, 1933-8007 online
DOI: 10.1080/08913811.2010.541686

facts into predictions about the future. If this view of expertise is correct, then it hardly seems surprising that Tetlock's careful, systematic analysis shows us what we should have expected all along: Experts are not required to be wise judges and they do not have any special competence or qualifications to foresee the future.

Facts combined with historical, social, cultural, political, and economic context are not a sufficient basis for scientific prediction and explanation. To be sure, facts, in the form of data, inform our estimation of variables thought to influence events, choices, and outcomes, but facts alone are rarely a good basis for prediction. The accumulated facts known to experts (or to anyone else) must be placed in the service of theory, transparent methods, and replicable analysis of evidence if we are to go from fact to inference.

Experts' Ideologies often Shape Their Predictions

What, for example, should we have made of the undisputed fact that both the United States and Soviet governments engaged in a massive accumulation of nuclear, chemical, and biological weapons during the Cold War? Liberals looked at the destructive power controlled by these two adversaries and applied intellectual constructs informed by the fear that the arms race would spiral out of control (Richardson 1960 and 1978; Jervis 1978), potentially leading to the destruction of life on our planet. They pressed for arms-control treaties, predicting that these would be the means to save the world from destruction, although there was scant evidence to show that arms-control treaties reduce the likelihood of war (Altfeld 1983; Koubi 1994). The evidence is stronger that arms-control agreements reduce the lethality of war (conditional on there being a war), because uncontrolled arms races increase the cost of war without a concomitant increase in expected benefits. Thus, more arms imply fewer wars but if war occurs, arms control implies more wars, but of lower average lethality.

Conservatives, in contrast, looked at the destructive power that had been amassed during the Cold War and, relying on their post-Munich *Realpolitik* intellectual constructs (Morgenthau 1973; Waltz 1979), concluded that vigilant deterrence shaped by a credible commitment to meet force with force would control the expansionist urges of adversaries. They predicted that improved weapons technology and military alliances would protect the world from destruction, noting a diminution in the incidence of war while tending to bypass evidence concerning the lethality of those wars that did occur.

Neither conservative nor liberal foreign-policy analysts lacked expert knowledge of facts, but the inferences they drew from a common set of facts were, in part, products of their political opinions and ideology, buttressed by the selective airing of historical examples (Rotberg and Rabb 1988). That each side knew facts added nothing to the transparency, "objectivity," or reliability with which they translated fact into prediction. Why anyone expected otherwise is a mystery for those of us, including of course Philip Tetlock, who believe in the merits of the scientific method over personal opinion or judgment.

Which of the many facts that are known to experts are germane to a problem ought to be determined by theory and prior evidence. Yet in identifying experts to survey in the hope of learning about their judgment, Tetlock's *Expert Political Judgment* (2005) allowed experts' sheer knowledge of large quantities of facts to dominate their knowledge of theory and methods as a selection criterion. What he has given us is a transparent, reproducible evaluation of the idiosyncratic, generally opaque means of assessment adopted by different experts rather than a systematic assessment of alternative methods for arriving at judgments about the future.

To be sure, there is a general propensity to believe in expert judgment. The Department of State and most of the "intelligence community," for instance, is organized around the perhaps-misguided assumption that expert knowledge translates into expert judgment. The ostensible reason to hope that experts' judgment is better than nonexperts' lies in the degree and substance of the experts' expertise—that is, their knowledge of many facts about the subject in which they specialize. It seems, for instance, that Tetlock has accepted the notion that someone who has built a career around studying China or studying relations between Israel and Palestine, for instance, should be better at predicting what will happen in their area of expertise than someone who has studied, for instance, decision making, negotiation, credible commitment, and strategic interaction.

Methods of Prediction

The persistent belief in expert political judgment is nowhere more surprising than in the arena of foreign policy. Here is an arena in which a blind eye is too often turned to methodical analysis in favor of expert judgment, despite the evidence in favor of methodical analysis. This is doubly surprising because the stakes riding on foreign policy choices are enormous. The social sciences have a demonstrated track record of better prediction than is attained through expert judgment in this and numerous other high-stakes arenas. Yet, as Stanley Feder (2002, 119), a former intelligence analyst, has hypothesized about a successful forecasting methodology introduced into the CIA, "this kind of systematic analysis does not fit into an organizational culture that sees an 'analyst' as someone who writes reports, often evaluating and summarizing available information. In contrast, people who use models and quantitative techniques are considered 'methodologists.'"

The social and behavioral sciences provide diverse approaches to successful prediction. Let's be clear that "successful" prediction does not mean perfect accuracy. The standards against which to evaluate prediction include, first, whether a particular method or approach outperforms available alternatives; then, whether the particular method or approach has been tested "out-of-sample"—that is, against outcomes not known at the time of the prediction; and finally, whether the particular method or approach can be taught or otherwise transmitted so that its success can be replicated by others. While these criteria surely are not exhaustive, they are a good start for measuring the value of predictive methods and perspectives. Consider a clear and well-known example.

Before the advent of high-speed computers, election prognosticators showed their mettle by predicting presidential (and other) election outcomes based on methods such as focusing on bellwether districts or interviewing a small number of key elites or hypothetically "representative" citizens. Although popular books are still written based on these sorts of methods, it is widely recognized today that they are not the best ways to predict election outcomes. Certainly all of the major media outlets long ago added to these soft methods more rigorous, reproducible analysis. Today's debate over the prediction of electoral outcomes revolves around the relative advantage of different types of survey techniques (national election surveys; various forms of stratified, weighted sampling; etc.), different regression-based statistical estimations of the impact of key economic indicators (local or national unemployment, inflation, economic growth rates, consumer confidence measures), and other methods rooted in the rules of vote aggregation, demographics, redistricting choices, and prediction markets. Each of these approaches outperforms the more colorful but less reliable methods solely relied upon in the past. Now, the pundit's expert judgment is largely informed by social-science theory and methods.

Statistical analysis has added greatly to the potential to sort out the relative importance of contending variables as instruments for advancing prediction. As such, it has become an integral part of academic, government, and business efforts to foresee the future by extrapolating from past patterns to out-of-sample circumstances. But as with any method, statistics has its limitations: not simply the ease with which people ignorant of statistical methods can be fooled by the framing of statistical results, but the inherent limitations of statistical methods in certain areas of predictive interest.

Statistics, simplistically speaking, are good at projecting the future based on underlying patterns in the past. As such, they are particularly good at relatively low-stakes issues like voting or market research where sample sizes are large. While the aggregate consequences of voting and market success are important for those seeking office or market share, the stakes for the individual voter or individual buyer generally are not very high. Moreover, statistics are not good at predicting fundamental breaks from the past. When earlier patterns between dependent and explanatory variables break down, statistical prediction tends to fare poorly. Such a breakdown seems especially likely when decision makers face high stakes; that is, high potential costs as well as high potential benefits from alternative choices.

The preferred substitute for statistical studies, especially among foreign-policy experts, is, not surprisingly, their forté: area expertise. After all, what could be more obvious than that if one wants to anticipate changes in, for instance, human-rights policy in China, then one must consult with someone who has expertise about China's culture, social norms and mores, as well as its political history and practices. Yet it is equally plausible that one ought to consult with someone whose expertise lies in the factors that lead government leaders to tighten or loosen controls over human rights. The former perspective emphasizes the assumption that there are no general covering laws addressing changes in human rights, but rather that idiosyncratic factors make decisions in China likely to be different from decisions in, say, India or Nigeria or Venezuela or Belgium.

But we have scant systematic evidence to support the perspective that area expertise translates into insight into predicting social change (as distinct from explaining known, past outcomes). In fact, Tetlock's meticulously assembled evidence contradicts this received wisdom. There is, in contrast, pretty solid evidence that insight into general patterns of change in respect for human rights provides a reliable means to predict or anticipate changes (or its absence) in China or anywhere else (Poe and Tate 1994; Poe, Tate, and Keith 1999; Bueno de Mesquita et al. 2005; Davenport 2007). Still, we can be confident that the newness of systematic research in this arena suggests that it will improve in the future. So far, it is likely that there are strong selection effects limiting the success of the covering-law (hedgehog) perspective. The relatively slim number of quantitative analysts in comparative and international politics suggests that we have barely given such covering-law approaches an equal opportunity to demonstrate their strengths and weaknesses. Thus, from an evidentiary perspective, the type of expertise that is relevant remains an open question. There is a propensity to associate area or country expertise with expert judgment, but Tetlock has persuasively shown that this propensity is flawed.

In my view, replicable theory that establishes causal or probabilistic links between facts and (especially, out-of-sample) results is, or ought to be, the foundation of scientific prediction. In the remainder of this essay I elaborate on that claim and show how the application of logic to evidence can help us create reliable predictions about important, non-obvious future events and developments. I do not merely assert this but offer a published record of evidence that reliable prediction of non-obvious political developments not only can be done, but has been done.

Game Theory as a Predictive Tool

Although many other excellent methods have been developed to assist in prediction (content analysis, spatial models, operations research, simultaneous equation systems, evolutionary models), my focus is on the application of game theory to the prediction of important, non-obvious policy outcomes. I believe that game-theoretic predictive models work better than expert judgment. I refer not only to game-theoretic models that I have used, but to excellent models developed by others, such as Thomson et al. 2006, Schneider et al. 2010, and models reported on in Schneider et al. 2011.

Tetlock criticizes game theory as a means to facilitate prediction. In his review of three recent books, including my *The Predictioneer's Game* (2009), he repeats the example used in *Expert Political Judgment* of a game in which people are asked to guess a number. He writes:

> Consider what happened when, many years ago, the *Financial Times* ran a guess-the-number competition for its readers (promising the winner an all-expenses-paid trip on the Concorde). The task was deceptively simple: predict a number between zero and one hundred such that your guess is as close as possible to two-thirds of the average guess of all other players. Some readers guessed 33 1/3—and were classified as strategically naive. They assumed that others would pick numbers randomly between zero and one hundred, which averages out to 50, and two-thirds of 50 is 33 1/ 3. These forecasters made the beginner's mistake of failing to factor in the incentives at work. Other readers guessed zero. They were too clever by half. They assumed that everyone knew as much game theory as they did—and they quickly reasoned through the deductive sequence: everyone knows that everyone knows that the first-order answer is 33 1/3, but if everyone converges on that answer then the correct answer is 22 1/6, and if everyone converges on that answer then the correct answer is two-thirds of 22 1/6, but if everyone converges on that answer... the theoretically correct answer—the Nash equilibrium—is zero. But the real answer for readers of the *Financial Times*, a pretty savvy group, was around 18, roughly halfway between the strategically naive answer of 33? and the too-clever answer of zero (Tetlock 2009).

Tetlock's use of this example to indict game theory strikes me as inappropriate. A good starting place in examining the FT's contest is to ask what game the participants were actually playing. There are at least two possible answers: (1) the game the FT hoped they would play; that is, choose the Nash equilibrium number; and (2) the game they probably played: that is, maximize their personal expected payoff from participating in the FT's contest. Are these the same problem? They are not.

Surely both the FT staff and participants in the contest could readily and reasonably assume there would be many ties for the winning answer.

In reality, participants were entering a lottery with a 1/X chance of winning a trip, with X equal to the number of "correct submissions," with correct being based on the modal answer and not the Nash equilibrium of the game the FT perhaps thought people were playing. Imagine, for instance, being a player who was "too clever." Such a player would anticipate that the Nash equilibrium answer is 0 and that if everyone proposed 0 then their expected value for the trip was just (1/X)*(Utility of the trip) with 1/X, the odds of winning the prize, presumably being very small. Indeed, the odds might have been sufficiently small that the strategically rational player would have submitted an answer larger than 0, hoping to distinguish herself from the crowd just enough to improve her expected value from the game, realizing that others too would try such a gambit by offering answers greater than 0. This sort of strategic bidding is a commonly observed pattern in prediction markets, where participants game, so to speak, the game. Tetlock has not considered how much more complicated this problem is than being asked for the Nash equilibrium of the simple game that the FT seemed to think people were being asked to play. That game, however, merely set the context for the expected payoff calculation against sophisticated players who would spread bids out above 0, much as one might do in a trembling-hand equilibrium, a quantal-response equilibrium, or a game with sufficient time discounting or other costs to limit the regress in the calculation to just 4 or 5 iterations.[1]

Confusion between "the right answer" and maximizing expected value from participation is a common white-noise problem in laboratory experiments. As Tetlock correctly anticipates, the better test of game theory's potential value is whether players in a game that actually has high stakes for them play close to the Nash equilibrium in the real world rather than in a white-noise-sensitive laboratory experiment in which we are uncertain which game subjects played. As we will see, the evidence shows that applied game-theory models have enjoyed documented, peer-reviewed success in the face of real-world, real-time prediction about high-stakes questions.

Tetlock partly anticipates these responses. Thus, he follows the FT example and associated claims by writing, "Bueno de Mesquita might reply that such complexities matter in contrived lab settings but the real-world proof of the forecasting pudding is in the eating—and he claims a delicious 90 percent hit rate based on [a] CIA report." He is right in anticipating this response from me. Unfortunately, he tries to deflect this response by not quite reporting the whole truth to his readers, as we will see. He presents what the proper standard of accuracy ought to be as he sees it: "Good social scientists are, however, Missourians: they insist on tasting the pudding themselves. The debate must unfold in peer-reviewed outlets—and there must be open, level-playing-field competition across approaches." I fully agree. At the same time, I cannot help but wonder why he chose to ignore the hundreds of predictions in peer-reviewed journal articles and books that I and others have made using my models and other models. These are results that anyone can consult and that generally were published (or accepted for publication) before the outcomes being predicted were known.

The Predictive Record

This failure to consult the peer-reviewed track record is doubly troubling because some of these publications have explicitly placed predictions from my models in direct competition with other models, including prospect theory to predict the implementation of the Northern Ireland Good Friday Agreement and the run-up to operation Desert Storm in 1991 (Bueno de Mesquita, McDermott, and Cope 2001; McDermott and Kugler 2001); several logrolling models to predict policy decisions within the European Union (Bueno de Mesquita and Stokman 1994; Thomson et al. 2006; Bueno de Mesquita 2009 and 2011); and institutional models and spatial models to predict decisions on the Maastricht Treaty, European Union Council decisions, and many other policy issues (Thomson et al. 2006). My forecasting model generally outperformed the alternatives even though most of the comparison articles include a prominent exponent of an alternative approach as author or co-author. These comparisons have stimulated improvements in my models and those developed by others. For instance, 1994 tests against European Union data showed an advantage for my particular model, but a more demanding, more extensive test in 2006 (Thomson et al.) yielded better results for alternative models. A 2011 paper revisits the 2006 data and shows that my most recent forecasting model (the Predictioneer's Game, hereafter PG) outperformed the large set of alternatives reported on by Thomson et al. in

2006. A recent paper by Schneider et al. (2010) introduces a newer model that may outperform PG in the European Union context. There is not sufficient evidence yet to reach a conclusion one way or the other. Of course, it is exactly such a competition of ideas and models that provide the basis for scientifically grounded progress in explaining and predicting policy choices.

Tetlock (2009) goes on to dismiss the record of success attributed to my models by the CIA, while still ignoring the peer-reviewed publications:

> Impressive though the numbers cited in Bueno de Mesquita's book are, they are also—without getting into nitty-gritty technical details—devoid of significance. A 90-percent hit rate is, for example, no great achievement for meteorologists predicting that it will not rain in Phoenix. And it is no big deal even to achieve a 100 percent hit rate of predicting X—no matter what X may be—if doing so comes at the cost of an equally high false-alarm rate. Anyone can predict every war from now until eternity by simply predicting war all the time.

The reader is left to infer that my particular forecasting models achieve high accuracy because they predict the obvious. Tetlock never actually says that they are used to predict incredibly easy things like "no rain in Phoenix." Nor does he demonstrate that success is achieved by having a high false positive rate; he leaves that, as well, for the reader to infer, but mistakenly, as I demonstrate below.

Tetlock's complaint is neatly and clearly summed up when he states,

> Reading these three books, it is easy to feel like a frustrated shopper wandering aimlessly down the forecasting aisle in the supermarket of ideas. The products on offer are packaged well—but we have no objective benchmarks, no trusted *Consumer Reports*, against which to gauge performance. We have no idea whether we would be better off paying one of these consultancies gobs of money for their proprietary forecasts or simply downloading the latest odds from a high-profile prediction market that culls individual bets on world events such as Tradesport. Indeed, would we do as well relying on the dart-throwing chimps or mindless extrapolation rules, like "Predict the most recent rate of change"? So, caveat emptor. (Tetlock 2009, 59)

Thus runs Tetlock's indictment. But what is the actual evidence for my forecasting models? I begin by examining some of the published literature in academic outlets on this question. They are a start in the direction of a

Consumer Reports assessment but they are neither the end of the testimonial evidence, nor are they the hard evidence available to any reader who is prepared to take the time to go through the published, peer-reviewed record. Following this review I turn to some of that hard evidence.

Stanley Feder (1995 and 2002), who spent twenty years as a political analyst and as a National Intelligence Officer at the CIA and the National Intelligence Council, provides a sample list of issues to which my forecasting model was applied. James Ray and Bruce Russett (1996) augment that list with information they gathered in evaluating my forecasting models in an article in the highly regarded, peer-reviewed *British Journal of Political Science*. Combining the two lists, here is a sample of issues addressed by the models in question for the reader to examine and assess with regard to whether they are of the type, ''Will it rain in Phoenix tomorrow?''[2]

> (1) What policy is Egypt likely to adopt toward Israel? (2) How fully will France participate in the Strategic Defense Initiative? (3) What is the Philippines likely to do about U.S. military bases? (4) What policy will Beijing adopt toward Taiwan's role in the Asian Development Bank? (5) What stand will Pakistan take on the Soviet occupation of Afghanistan? (6) How much is Mozambique likely to accommodate the West? (7) How much support is South Yemen likely to give to the insurgency in North Yemen? (8) What is the South Korean government likely to do about large-scale demonstrations? (9) What will Japan's foreign trade policy look like? (10) What stand will the Mexican government take on official corruption? (11) When will the presidential election be held in Brazil? (12) Can the Italian government be brought down over the wage-indexing issue? (13) Who will succeed Ayatollah Khomeini in Iran? (14) Will China face significant internal political instability in 1989? (15) Who will win the Nicaraguan election (in 1990)? (16) Will the two Koreas be admitted to the United Nations? (17) What will be the price of (West Texas Intermediate light Sweet) crude oil? (18) Will the coup in the Soviet Union succeed or fail?[3]

Numerous peer-reviewed forecasts are identified by Ray and Russett in their footnotes.[4] Both before and since their article was published, there have been many more such predictions, by myself and others. The most recent peer-reviewed study applying my latest forecasting model to an important current event is Mousavi and Shefrin (forthcoming), so the list continues to grow, providing ample opportunity to test the false-alarm rates.

As should be evident from the sample of issues listed by Feder or Ray and Russett (or others; e.g., Allas and Georgiades 2001; Dixon and Nicoll 1999) and from the sample of subjects in peer-reviewed articles using my forecasting models, these models are routinely applied to difficult, often controversial policy problems. Whether examining government applications, private-sector applications, and peer-reviewed applications, forecasts generally examine politically (or economically) significant issues surrounded by substantial *ex ante* uncertainty. Hundreds of forecasts in peer-reviewed outlets have been accepted or published before the outcomes were known, and they are available for scrutiny.

Feder's CIA study, dismissed by Tetlock, goes on to report on the effectiveness of these forecasting tools in doing contingency analysis, scenario testing, and sensitivity testing. As Feder (1995) writes:

> For policy and intelligence agencies, one advantage of these models is that their data inputs are the observations of country or issue experts. Use of the models, particularly the way in which analysts provided the data, also made it easy to avoid analytic traps such as expecting the future to look like the past and failing to consider alternative outcomes. . . . In addition to providing policy forecasts, the models also make possible reliable inferences about the stability of a government and the emergence of new leaders. Within a parliamentary system, if a policy supported by the head of government or ruling party is defeated, the government collapses. When using Bueno de Mesquita's models, if the forecasted outcome is politically far from the position of the head of government, that leader is vulnerable to defeat. Several times in the past 20 years, we foresaw the collapse of a number of governments based on this kind of analysis. . . . Bueno de Mesquita's models facilitate surveying the future by making it easy to explore the implications of possible changes in a political environment. . . . Because the forecasts are conditional and conditions can change, the sensitivity analysis provides a list of political factors to monitor. Testing "what if" scenarios also provides an indication of how much change is possible and how quickly it can occur.

Models Beat Experts, But So Do Chimps. . . .

Tetlock's final challenge is a crucial one and is deserving of close and careful attention:

> I do not mean here to trivialize Bueno de Mesquita's predictive track record: he is rare among social scientists in keeping score and his performance is impressive—even to one as jaded as I. That said, I have reservations. It is unclear how Bueno de Mesquita would counter the argument that outperforming the individual experts is no grand feat. As already noted, it is not hard to beat individual experts in the forecasting game. From this perspective, Bueno de Mesquita's model may be accomplishing no more than what averaging routinely does—and even dart-throwing chimps can occasionally pull off. (Tetlock 2009, 66)

This, rather than the suggestion that my models are used to predict the obvious, is the telling challenge and it is one to which I happily respond. The counter-argument is two-fold. First, however regrettable it is that experts remain the standard to beat in the minds of non-academics and many social scientists, we should not dismiss beating the experts as substantively inconsequential. In fact, in a critique of my forecasting record, Kesten Green (2002) contends that the right measure is whether the model outperforms the experts. Unfortunately, he concludes, based on a selective report of Feder's 1995 findings, that my original model and the experts do equally well, without reporting Feder's 1995 assessment that my original model (which Feder calls Policon) hit the bull's eye—that is, was spot on—about 60 percent of the time, and that the experts who provided the data only hit the bull's eye half as often. They were in the neighborhood of the right outcome—on target, in Feder's terms—but they were not nearly as accurate. Thus, both the experts and the model were pointing in the right direction 90 percent of the time, but the model greatly outperformed the experts in precision (lower error variance), according to Feder. Feder also notes that in the cases he examined, when the model and the experts disagreed, the model proved to be right, not the experts (who were the only source of inputs for the model).

Tetlock and I agree that these models have beaten the experts most of the time. We also firmly agree that the best means of evaluating forecasting reliability is to test against alternative approaches. I already reported on several such tests in peer-reviewed outlets. Now I want to present tests against Tetlock's contention that "Bueno de Mesquita's model may be accomplishing no more than what averaging routinely does—and even dart-throwing chimps can occasionally pull off." Let's look at some hard evidence. I start with out-of-sample evidence that is still unfolding and then turn to past studies and the evidence for or against their accuracy.

The last two chapters of my 2009 book, *The Predictioneer's Game,* contain predictions about likely developments pertaining to politics in Pakistan during 2010 and 2011, predictions regarding relations between Iran and Iraq, and predictions about internal political developments in each of those countries between the summer of 2010 and the beginning of 2012. Those predictions are contingent on troop-commitment decisions by the Obama administration. The final chapter also made predictions about the Copenhagen summit. I have reviewed the accuracy of the Pakistan and Copenhagen predictions in an epilogue added to the paperback edition (written in February, 2010) and invited readers to do their own assessments.

The detailed, contingent developments predicted for Pakistan regarding its pursuit of militants from the Taliban and al Qaeda have proven to be quite accurate. They were neither obvious nor without controversy at the time they were made.

Of course, none of us know, at the time I am writing this article, what will happen between Iran and Iraq through 2012. I certainly urge readers to evaluate those predictions after events have unfolded.

There is no reason to limit such an evaluation either to a simple right/wrong dichotomy, or merely to a comparison against expert judgments about these events made roughly at the time I was writing the 2009 book.[5] As for Copenhagen, the predicted failure of that summit ran against the mainstream point of view at the time I wrote *The Predictioneer's Game*. Not only did the summit fail but it did so along the dimensions laid out through the model-based analysis. Thus, the *ex ante* predictions whose *ex post* realizations are now known proved reliable.

Predicting Democratization or Autocratization: 1981–2008

In an earlier book, *Predicting Politics* (2002, 153-55), I provided data for my model on 101 countries. The variable of interest was how democratic or autocratic the countries were and would become. The model in that book relied on three variables for each of the stakeholders: how much potential influence they could exert on the issue of interest; their then current position on the issue; and their salience or degree of focus on the issue. The PG model introduced in my 2009 book added a variable that measures the degree to which each player values reaching agreement with others or values sticking to their position on an issue even if it means defeat. That variable is called flexibility.

The data used in my 2002 book (explained below) to predict changes in democracy across countries were rather crude. Furthermore, the data were flawed in that they focused *only* on international pressure regarding governance, entirely ignoring domestic political pressures (and lacking a nuanced view of individual internal players), although these are of obvious importance. This is unfortunate, but it was the best that could be done with the data then available to me. The models could have used domestic-politics inputs, so the limitation in the analyses that follow is purely data availability and not a function of the models themselves. The data in the 2002 analysis and in the analysis to follow (based on PG) have not been updated by any developments after 1980 within or between the countries studied.

I used Polity's Democracy-Autocracy scale, normalized to 0 to 100 (originally −10 to +10) as the indicator of each government's 1980 position on democracy or autocracy. Higher values indicate more democracy and lower values more autocratic governance. Changes in the predicted value of this variable, year by year from 1981 to 2008 for each of about 100 countries, constitute the indicator of interest for assessing predictive accuracy. That is, each "bargaining round" of the models was equivalent to one year of calibration to match Polity's annual rescoring of countries on their autonomy-democracy variable. Thus, the predicted values for each year from 1981–2008 (the latest year for which Polity has reported data) are evaluated in comparison to Polity's *ex post* record of each country's score in each of those years.

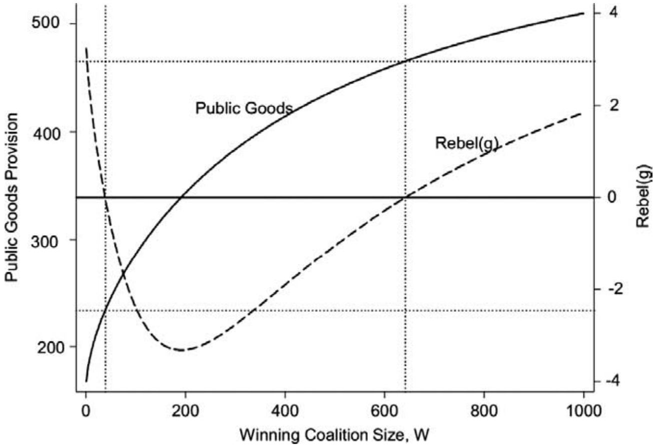
The measurement of potential political influence was based on the World Bank's 1980 GDP data. Although the model in my 2009 book *predicts* changes in each variable's value across each bargaining round and uses those predicted values, it is important to keep in mind that the only data fed into the PG model were 1980 data.

The Salience variable is intended to estimate how high a priority the issue of interest is to the players in the model. Salience was measured following the procedure explained in my 2002 volume (105–6). These estimates relied on ideas from the selectorate theory of politics (Bueno de Mesquita et al. 2003), in which it is assumed that leaders first and foremost want to survive politically. The method estimated the risk to a leader's political survival that was contingent on how long the incumbent had already been in office, the regime's Polity score, and other factors. As with the other variables, no data were brought into the model beyond what was known in 1980.

I have augmented the data on influence, position and salience by adding the variable required by PG; namely, the extent to which government leaders wish to come to agreement with others or are resolved to stick to their narrowly defined interests. This Flexibility variable is estimated based on theoretical results in Bueno de Mesquita and Smith's (2009) extension of the selectorate theory (Bueno de Mesquita et al. 2003), which is designed to provide a formal, game-theoretic explanation (and prediction) of endogenous institution change. From it is derived the functional form of conditions that encourage leaders to keep their current form of government (likely for those at the extreme ends of the democracy-autocracy issue space) or to lean toward change in either direction and to varying degrees (becoming more democratic or more autocratic).

Figure 1 illustrates the theoretically derived proposition regarding the inclination to become more democratic, more autocratic, or to remain the same. As the Flexibility data in the Appendix show, the functional

Figure 1. Endogenous Institution Change to a Smaller Winning Coalition or to a Larger Coalition.



Smaller W = more autocratic, larger W = more democratic
Source: Bueno de Mesquita and Smith 2009.

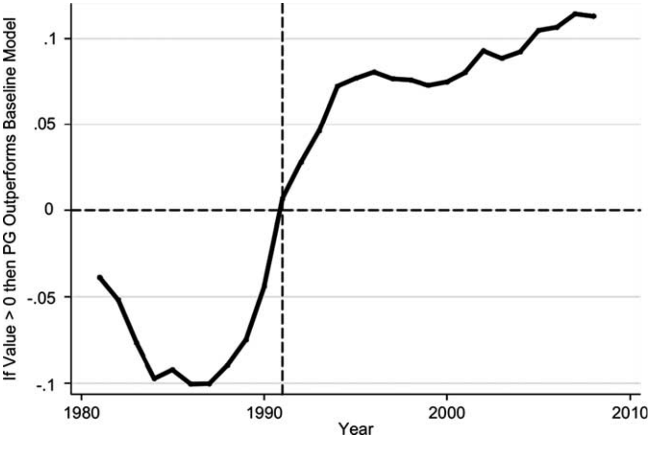
form in Figure 1 was implemented in a conservative way by assuming that regimes with no more than the minimum Polity score (0) satisfied Figure 1's condition for falling below the dotted vertical line on the left of the figure. This line shows the cut-point consistent with the theory in Bueno de Mesquita and Smith (2009), which indicates that governments below some minimal coalition-size threshold provide so few public goods that their adversaries cannot coordinate a credible mass movement or credible threat of rebellion that would jeopardize the regime and its institutions. Therefore, leaders of regimes that fall below that threshold do not need to counter threats of rebellion either by expanding or contracting their dependence on coalition support and their provision of public (g) goods. Likewise, only regimes with the highest Polity score (100) were assumed to satisfy Figure 1's condition for falling above the cut-point depicted as the vertical dotted line just to the right of the hypothetical coalition size of 600. Regimes above some (calculable) threshold provide so many public goods to their citizenry that they do not face a credible threat of rebellion against the regime and institutions of government because citizens are sufficiently satisfied with their well-being that they do not care to rebel. Thus, regimes above the upper-bound threshold are not expected to alter their governance institutions. Regimes that fall between these two thresholds are likely to become more autocratic (reduce public-goods provision) or become more democratic (expand their provision of public goods), with the flexibility to make a choice one way or the other (to be predicted by PG) following the inverse of the rebel function plotted in Figure 1. That is, regimes are most flexible when they approach the minimum of the Rebel (g) function in Figure 1 and less flexible about contracting or expanding democracy as they move farther away from that minimum.

The form of governance does not typically change rapidly or markedly. Thus a best first-cut, naïve, baseline prediction of regime change is that a regime's form of governance will remain the same. I tested this baseline model against the Predictioneer's Game model. Following Tetlock's reasonable contention that for many problems, today's answer is the best predictor of tomorrow's answer, I have relied on the 1980 Polity data as the baseline for predicting values in each subsequent year. In fact, Ian Budge and Dennis J. Farlie (1981), in their effort to predict regime change through 1980, find—consistent with Tetlock's supposition—overwhelmingly and not surprisingly that the best predictor of a regime's type is the regime type the state had in the previous time period. Thus, I have constructed what I believe would have been a level playing field to compare the baseline prediction (today is the best predictor of tomorrow) against PG's predictions for anyone interested in regime changes back in 1980.

Since regimes change slowly, Polity's scores for 1980 ought to be a good predictor for Polity scores in 1981, a bit less good for 1982 scores, and so forth. If PG is no better than throwing darts or simple averaging, then PG ought not to do as well using the same 1980s data as is done by just predicting constancy in the Polity democracy-autocracy scores.

Figure 2 plots the difference in the predicted errors for the baseline model minus PG year by year across the set of countries displayed in the Appendix.[6] For Figure 2, I simply took the mean of the absolute deviation of the "true" Polity score and subtracted from it the mean of the predicted score each year for PG and for the Polity scores in 1980. The results are, I believe, instructive. Between 1981 and 1990, the baseline prediction that the regime score in 1980 is the best predictor of the regime score in subsequent years is supported. Indeed, for the first few years of the decade of the 1980s, the PG forecasts were getting ahead of the world, predicting change before it happened (beginning dramatically with the end of the Cold War in 1989–1991). But once we get past the first few years of the 1980s, the baseline result and the modeling results first begin to converge toward an error difference of 0,

Figure 2. Difference in Predictive Errors, PG versus Baseline Model



and then the forecasting model equals and overtakes the baseline as the better forecaster. That is, the real world catches up to the Predictioneer's Game model by about 1990, and then the PG forecasts match the real world better than the baseline predictions do. Indeed, the farther out we get on this slow-moving variable measuring democratization, the better the PG model does compared to the baseline notion that today is the best predictor of tomorrow. Yet neither the baseline predictions nor PG's predictions had access to any information not known in 1980.

We can see the information from Figure 2 in more detail in Table 1. Here I report the median error, mean error, and standard deviation for three time periods: 1981–1990; 1991–2000; and 2001–2008. Errors are computed as the absolute value of the difference between the Polity Score reported for each country each year and the predicted Polity score for each country each year, based on the Predictioneer's Game model (PG) or the Baseline (1980 Polity Score), with this quantity divided by 100 (the maximum possible error) to yield a percentage error.

Each of the three reported values associated with error tell an important part of the forecasting story. The median error minimizes the impact of extreme outliers while the mean gives an overall sense of how well each perspective does. The standard deviation provides further information on the confidence we can attach to the predictions. The "best" performer on each dimension is shown in italicized lettering in Table 1.

Table 1, of course, is just another, more detailed way to look at the results seen visually in Figure 2. But the Table's added details are informative. Although the baseline model had a smaller median and mean error than did PG in the decade of the 1980s, PG's errors are still fairly small and its standard deviation is very slightly better than for the baseline model. So PG is not a bad predictor, albeit it got a bit ahead of changes in the world in the near term; that is, up through 1990. After 1990, PG outperforms the baseline predictions on almost every indicator and almost every year. PG's mean errors become substantially smaller than those for the baseline model, and so do the standard deviations of the errors.

We can look at these results still another way. The study by Budge and Farlie (1981), cited earlier, was a statistical attempt to predict regime changes based on national characteristics, including such indicators as GDP, population, working population, number of newspapers per thousand population, and population per physician (ibid., 339). They defined regime types in several ways that are highly correlated with the

Table 1. Forecasting Errors: The Predictioneer's Game Compared to the Baseline

	1981–1990			1991–2000			2001–2008		
	988			964			768		
Observations	Median Error	Mean Error	Std. Dev.	Median Error	Mean Error	Std. Dev.	Median Error	Mean Error	Std. Dev.
PG	0.09	0.16	*0.19*	0.15	*0.20*	*0.21*	*0.12*	*0.19*	*0.21*
Baseline (1980)	*0.00*	*0.08*	0.20	*0.10*	0.26	0.30	0.15	0.29	0.31

Italicized numerals correspond to the best performer in each dimension.

Table 2a. Success of Statistical Model at Predicting Regime Change, 1950–1970:

Actual →Predicted ↓	Change	No Change
Change (Budge and Farlie)	18	27
No Change	23	70

Polity data, including the number of political parties, competitiveness in selecting leaders, and the like. They used the data to predict regime changes in five-year swaths. Their false positive and false negative rates were sufficiently high, as seen in Table 2a, that one would have done no better with their results than by predicting no regime change in all cases.

Table 2b provides a useful contrast to Budge and Farlie's statistical approach. Here, a forecaster relying on PG would make 69 percent fewer errors by predicting whether a given regime would change or not based on the game-theoretic results instead of by simply predicting the *ex post* known modal category. Whereas Budge and Farlie's early statistical effort suffers from many false positives (27 predicted changes when none occurred) and many false negatives (23 predicted absences of change when regime change occurred), equaling 36 percent of all observations, PG has only 6 false positives and 2 false negatives out of 98 predictions, equaling only 8 percent of the observations. This is not to criticize Budge and Farlie who, after all, undertook a pioneering cross-national study that required them to invent statistical methods to handle some of the issues in the data that were available to them. Rather, comparing PG to the most comparable known undertaking based on statistics highlights the minimal amount of false positives and false negatives in PG's predictions.

Continuing with analysis based on distinctions made by Budge and Farlie, Table 3 looks at predicted values and actual values for all of the

Table 2b. Success of Predictioneer's Game at Predicting Regime Change, 1980–2008:

Actual →Predicted ↓	Change	No Change
Change	70	6
No Change	2	20

Table 3. PG's Success at Predicting Autocracy or Democracy

Actual → Predicted ↓	Most Autocratic	Transitional	Most Democratic
Most Autocratic	109	63	36
Transitional	58	1263	604
Most Democratic	0	1	642

country-years in my data from 1981 through 2008 according to whether, in each country-year, a regime was extremely autocratic (normalized Polity scores from 0 through 15); transitional or moderately autocratic (scores above 15 and below 85); or democratic (scores of 85 to 100). This tougher standard sorts predicted and actual regime types year-by-year into politically significantly different categories.

A naïve predictor, who always predicts that a regime falls in the intermediate category, would get 48 percent of the predictions right. Using PG's predicted categorization of country-years gets 73 percent of the cases right, despite the reliance on rather crude estimates of PG's variables.

When predicting a slow-moving variable like regime type, given a reasonable amount of time, PG substantially outperforms the baseline model and may—we will have to await more current studies than Budge and Farlie's—outperform statistical efforts as well. PG's defect in this instance seems to be that it got out ahead of the sudden and dramatic regime changes following the 1989–1991 end of the Cold War, but not by that much. For long-term planning purposes, PG would have been preferable to the baseline model.

European Union Predictions

A second set of comparisons should further help solidify confidence that forecasts that rely on expert judgment for inputs but modeling logic for outputs can provide reliable, replicable insights into future developments. To evaluate this possibility I turn to results from Bueno de Mesquita 2011 that compare the performance of the PG model to the best alternative perspective derived from a detailed study of decisions in the European Union (Thomson et al. 2006).

Robert Thomson and Frans Stokman kindly provided me with two EU data sets with which to test the Predictioneer's Game. Elsewhere I have reported more extensively on these tests. Here I focus just on two EU data sets that approximate the conditions for which the Predictioneer's Game are suited; that is, iterated rather than repeated games.

One data set consists of 9 issues from the European Union. These data were provided to me by Thomson and represent the only EU data used here for which there are expert estimates of the flexibility variable that PG requires.

A second set of tests parse Thomson et al.'s (2006) collection of 162 EU issues, which unfortunately do not include estimates of the flexibility variable. Thirty-seven of these issues have *ex ante* conditions that come close to the iterated non-cooperative game environment assumed by PG. Unlike the remainder of Thomson et al.'s 162 issues, these 37 had no recursion values. This means that 125 of the 162 issues had been discussed before in the EU, so the EU had an established position on them. Given the highly cooperative nature of the repeated-play EU decision-making environment, these 125 issues are particularly likely to deviate from the analytic context for which PG—a non-cooperative, iterated game model—was designed. I exclude these 125 issues here (but see Bueno de Mesquita 2011 for their analysis), as they are better suited to models, like those designed by Frans Stokman and his colleagues, that focus on repeated games and logrolls. The 37 issues without recursion points were more likely to involve real negotiation and exertion of leverage, since there was not a prior policy to which the European Union members had agreed and knew they could revert. Thus, while not as good a test as the 9 issues for which I have data on flexibility as well as position, salience, and potential influence, these 37 cases still at least have a heightened probability of being contextually appropriate for PG. Because I have no basis for choosing variation in the flexibility variable for these 37 cases, I set each European Union member's initial value at 50, in the middle of the flexibility scale. As we will see when examining the 9 issues for which I have flexibility data, setting the value at 50 for everyone introduces a considerable increase in predictive error (as should be expected).

As with the previous analysis, I am interested in how the models perform in terms of the absolute mean percentage error, the median percentage error, and the standard deviation of the error. Christopher

Achen (2006) reports that the weighted mean position of the initial EU data does about as well as or better than any of the strategic, institutional, logrolling, or other models tested in Thomson et al. 2006.[7] He suggests that the added complexity of models that make assumptions about player interactions, institutional constraints, etc., does not yield enough of an advantage to warrant using them. That is, he comes down firmly in Tetlock's camp in holding that forecasting models must be evaluated against alternatives, and he finds that simple averaging seems to do rather well. Of course, the underlying theories behind the models that Achen rejects dictate the use of influence and salience, so even the initial weighted mean value that he uses is informed by theory; but still, Achen's point is an important one.

The principles of parsimony and Occam's razor remind us that we have no need for complex algorithms if we can do as well with a simple approach. Below, I report results based on Achen's best-performing instrument for the EU data; that is, the weighted mean voter position. I also report the weighted median voter position, as this is another prominent basis for prediction among political-science modelers. As we will see, however, while Achen's finding may hold for the EU in general, it does not hold when compared to PG either in the case of the 9 issues for which I have complete data or the 37 issues for which I lack flexibility variance. That is, as the analysis focuses on the more competitive, non-cooperative subsets of the EU data (i.e., the subsets that represent a more appropriate test of the theory behind PG), the added complexity introduced by PG appears to be warranted.

Table 4 displays the error rates across the 9 issues for which I have complete information. As can be seen, there are substantial differences in the performance of the models. The Predictioneer's Game model is by far the best fitting, whether goodness of fit is assessed in terms of median

Table 4. Model Error Rates When Data Include a "Resolve" Variable

Model (Abs Error)	Median	Mean	Std. deviation	no. of cases
Predictioneer's Game	7.7	8.9	8.1	9/Thomson
Median Voter Prediction	20.0	29.4	33.7	9/Thomson
Mean Voter Prediction	12.5	11.8	9.8	9/Thomson

or mean error. Not only are the errors small, but so too is the standard deviation, reflecting a tight fit with the actual outcomes across these 9 issues. We can find some support for Achen's observation that the initial weighted mean voter position is a good predictor. It clearly outperforms the median-voter prediction. However, here, where there are complete data for PG, the initial predicted mean (or median) positions based only on input data with no strategic interplay fare poorly compared to PG. PG's weighted median error is about 50 percent smaller than the initial weighted mean voter prediction, and about one-third the initial median voter prediction.

Table 5 helps set the stage for the remaining analysis. As I have emphasized, beyond the data on the 9 issues provided to me by Thomson, I do not have an estimate of PG's flexibility variable, so I set its initial value arbitrarily at 50. Because this undoubtedly introduces error, it is important to keep in mind. How large is that error? We can approximate the predictive error introduced by data limitations by substituting 50 for the actual values for flexibility for the 9 Thomson issues. Table 5 shows the comparative results with and without expert-estimated variance on flexibility.

Table 5 shows a dramatic difference in goodness of fit. Whereas the median error was only 7.7 percent with complete data for PG, the median error rises to 10.1 percent and the mean error rises from 8.9 percent to 21.9 percent without proper inputs on this variable. The same is true for the variance in predictive error. With complete data, the standard deviation of the predictive error is around 8. Without variance on the flexibility variable, the standard deviation is 31. Thus we

Table 5. Comparing Results with Complete and Incomplete Data on Flexibility

Models: Comparing Complete Thomson Data To No Flexibility Data	Median	Mean	Std. deviation	no. of cases
Predictioneer's Game, With Flexibility Data	7.7	8.9	8.1	9/Thomson
Predictioneer's Game, Flexibility = 50	10.1	21.9	30.8	9/Thomson

can see that there is a substantial degradation in the predictive reliability of the model when the data do not provide information on the flexibility variable.

Bearing in mind the significant increase in predictive error introduced by not having information on the values for the flexibility variable, I now turn to the test of PG for the 37 EU issues without a recursion point, drawn from the set of 162 issues used by Thomson et al. 2006, Schneider et al. 2010, and others.

Table 6 again offers encouragement for the belief that PG, despite the handicap of a significant artificially induced measurement error, outperforms its most successful alternatives. PG's median error is smaller than the median error for the Mean Voter, but it is not as small as the median error based on the Median-Voter model. While this points toward the initial weighted median-voter position as a valuable prediction, we should note that PG's mean percentage error of prediction is notably better than the mean predictive error for the Median- or Mean-Voter models, despite the absence of data on the flexibility variable. Furthermore, the standard deviation of the predictive errors of PG is smaller than for either other model, again despite the introduction of known measurement error only into the PG prediction.

Table 7 examines the errors of prediction across the models for the entire 162 cases used in Thomson et al. 2006. Here, even with maximal impact of measurement error and with PG misfit to a largely cooperative database, PG proves advantageous when viewed from the perspective of its median percentage error (12.7 percent) compared to the initial median error based on the Mean-Voter prediction without strategic interplay

Table 6. No Flexibility Data, Issues without a Recursion Point; Likely to Be Less Cooperative

Model (Abs Error)	Median	Mean	Std. deviation	no. of cases
Predictioneer's Game	8.2	16.9	24.8	37/No Recursion
Median Voter Prediction	5.0	19.8	29.8	37/No Recursion
Mean Voter Prediction	8.6	19.4	28.0	37/No Recursion

Table 7. Tests with Maximal Measurement Error for the New Model

Model (Abs Error)	Median	Mean	Std. deviation	no. of cases
Predictioneer's Game	12.7	22.8	25.9	162/EU Decides
Median Voter Predictions	20.0	28.2	30.7	162/EU Decides
Mean Voter Predictions	14.4	22.5	25.5	162/EU Decides

(14.4 percent). The mean predicted errors and the standard deviations of the errors are essentially the same.

* * *

Philip Tetlock's *Expert Political Judgment* is a powerful, systematic, compelling indictment of expert judgment as the means to predict and plan for the future. Yet his "Missourian" reserve with regard to the Predictioneer's Game model seems overstated if not completely misplaced. Real-time predictions published in peer-reviewed journals of my older forecasting model (which is systematically out-performed by PG [Bueno de Mesquita 2011]) show that it does better than experts, including the experts who provide the data inputs it uses. Real-time forecasts in numerous publications also show that PG and my older model have regularly outperformed competing, alternative models, such as prospect theory, logrolling models, mean-voter models, median-voter models, and many others.

I have provided testimonial evidence by academics and government experts regarding the performance of my original forecasting model and its variants. I have also provided statistical evidence that the Predictioneer's Game outperforms those models to which it has thus far been compared, including power-index models, logrolling models, institutional models, the mean and median voter models, and my earlier forecasting models. Of course, more such tests against more models in blind test settings are desirable and necessary. On that both Tetlock and I can most assuredly agree.

Appendix: Data Used to Predict Changes in Polity Democracy-Autocracy Scores, 1980–2008

Country	Influence	Position	Salience	Flexibility
Algeria	3.16	5	86.2	12.5
Argentina	11.27	5	60.22	12.5
Australia	11.28	100	28.85	0
Austria	4.87	100	20.92	0
Belgium	6.71	100	65.22	0
Benin	0.24	15	45.27	32.5
Bangladesh	5.77	30	65.48	100
Bolivia	0.68	15	99.9	32.5
Botswana	0.11	80	21.91	30
Brazil	32	30	83.79	100
Burundi	0.12	15	61.45	32.5
Bulgaria	2.13	15	28.23	32.5
Burma	1.04	10	29.81	22.5
Canada	20.84	100	99.9	0
Cameroon	0.64	10	27.86	22.5
CAR	0.1	15	85.95	32.5
Chile	2.66	15	47.07	32.5
China	58.49	15	61.24	32.5
Colombia	4.79	90	31.66	16
Comoros	0.01	25	74.5	77.5
Congo Brazzaville	0.19	10	96.03	22.5
Costa Rica	0.52	100	58.26	0
Cyprus	0.2	100	47.28	0
Czechoslovakia	3.49	15	54.9	32.5
Denmark	3.56	100	30.18	0
Dominican Republic	0.82	80	64.71	30
Ecuador	1.61	95	74.67	9
Egypt	4.12	20	38.85	52.5
Ethiopia	0.74	15	67.8	32.5
Fiji	0.14	95	26.51	9
Finland	3.18	100	13.35	0
France	38.84	90	29.85	16
GDR	7.84	5	41.24	12.5
Germany	45	100	26.59	0
Ghana	0.64	80	78.92	30
Guinea Bissau	0.02	15	48.57	32.5
Greece	3.49	90	34.9	16
Guatemala	1.09	25	73.76	77.5
Guyana	0.09	15	33.31	32.5
Haiti	0.34	5	43.98	12.5

Appendix (*Continued*)

Country	Influence	Position	Salience	Flexibility	Country	Influence	Position	Salience	Flexibility
Honduras	0.34	55	71.52	65	New Zealand	1.98	100	31.24	0
Hungary	3.28	15	23.63	32.5	Nigeria	6.27	85	77.84	23
Iceland	0.16	100	64.59	0	Niger	0.24	15	52.02	32.5
India	37.18	90	99.9	16	Norway	3.05	100	33.25	0
Ireland	1.42	100	70.25	0	Netherlands	9.79	100	40.14	0
Iran	8.24	40	84.91	86	Pakistan	5.62	15	67.49	32.5
Israel	1.88	95	46.2	9	Paraguay	0.49	10	23.62	22.5
Italy	35.73	100	99.9	0	Peru	3.05	85	99.9	23
Jamaica	0.31	100	99.9	0	Philippines	5.57	5	33.15	12.5
Japan	72.13	100	99.9	0	Papua	0.33	70	39.35	44
Liberia	0.11	15	99.9	32.5	Poland	9.64	20	38.34	52.5
Lesotho	0.08	15	32.36	32.5	Portugal	2.98	95	42.73	9
Luxembourg	0.27	100	64.11	0	Rhodesia	0.52	75	99.9	37
Mauritania	0.08	15	99.9	32.5	Korea	7.23	10	5.77	22.5
Malaysia	3.21	70	45.19	44	Romania	1.94	10	32.76	22.5
Mauritius	0.24	95	20.13	9	South Africa	6.33	70	66.09	44
Mexico	24.89	35	55.68	93	Saudi Arabia	7.9	0	52.34	5
Mali	0.21	15	36.83	32.5	Senegal	0.39	40	24.95	86
Mozambique	0.69	10	56.84	22.5	Sierra Leone	0.23	15	42.65	32.5

Appendix (*Continued*)

Country	Influence	Position	Salience	Flexibility	Country	Influence	Position	Salience	Flexibility
Singapore	0.99	40	22.39	86	Trinidad	0.75	90	14.32	16
Somalia	0.26	15	38.52	32.5	Tunisia	0.99	5	25.78	12.5
Spain	16.94	95	41.96	9	Turkey	7.83	25	84.64	77.5
Sri Lanka	1.48	80	57.18	30	UK	35.12	100	66.47	0
Sudan	1.02	15	38.48	32.5	Upper Volta	0.2	15	33.84	32.5
Swaziland	0.11	0	34.6	5	Uruguay	0.91	15	59.41	32.5
Sweden	6.35	100	63.27	0	USA	213.63	100	34.17	0
Switzerland	5.54	100	99.9	0	Russia	99.64	15	29.43	32.5
Taiwan	4.87	15	74.26	32.5	Venezuela	6.75	95	32.32	9
Tanzania	0.53	20	29.33	52.5	Yugoslavia	7.61	25	18.34	77.5
Thailand	6.24	60	99.9	58	Zambia	0.34	5	32.25	12.5
Togo	0.12	15	35.16	32.5					

NOTES

1. Consider, for instance, what a player looking four or five iterations beyond the idea of randomly distributed "votes," with an average value of 50, would have submitted as the answer. Rather than doing the infinite regress—being, in Tetlock's terms, clever but not "too clever by half"—such a player might have deviated from the "too clever" answer of 0 and assumed that errors or sophisticated spreads would have produced a "winning" answer that was the average of only 4 or 5 regresses back from 50. If players, trying to get better odds of winning (that is, improving over 1/X), assumed others were doing the same and so randomized (playing mixed strategies) over the *average* of the first cut answer (2/3 of 50 = 33 1/3), and then each of the next 4 (2/3 of 33.333; 2/3★2/3★33.333; 2/3★2/3★2/3★33.333; and 2/3★2/3★2/3★2/3★33.333, then their answers would have been between 20.06 (4th regress) and 17.37 (5th regress), the average of which is 18.715; that is, just about the winning answer of 18. This example assumes a limit imposed on the potentially infinite regress leading to the Nash equilibrium value of 0, but it does illustrate how a clever but not "too clever by half" player could reasonably and strategically have ended up with 18 as the answer.
2. These questions were studied during specific time periods, of course, and made predictions for specific periods in the future. These details are left out here and in the Feder and the Ray and Russett studies. For the studies in peer-reviewed outlets the reader can check on these additional details.
3. The years of these predictions are classified. The CIA has declassified the fact that these questions were asked, but not *when* they were asked.
4. For example, Beck and Bueno de Mesquita 1985, 103–122,; Bueno de Mesquita 1990; Bueno de Mesquita and Iusi-Scarborough 1988; Bueno de Mesquita and Kim 1991; Bueno de Mesquita and Organski [1990] 1990; James 1998; Kugler 1987; Morrow, Bueno de Mesquita, and Wu 1993; Newman and Bridges 1994; Organski and Bueno de Mesquita 1993; Wu and Bueno de Mesquita 1994; Bueno de Mesquita and Stokman 1994; Kugler, Snider, and Longwell 1994.
5. The finished manuscript was submitted in the late summer of 2008 and the copyedited manuscript was ready for the printer around the beginning of summer, 2009.
6. There is some attrition in cases, though not very much, as several countries ceased to exist between 1980 and 2008, a question I did not address in these forecasts. I treated Russia as the successor state to the Soviet Union, and Germany as the successor state to East and West Germany (counting them as one country after unification).
7. The absolute weighted mean percentage error is calculated as |Predicted – Observed| with the weighted mean outcome computed as

$$\frac{\sum_{i=1}^{n}(\mathit{Influence}_i)(\mathit{Salience}_i)(\mathit{Position}_i)}{\sum_{i=1}^{n}(\mathit{Influence}_i)(\mathit{Salience}_i)}$$

REFERENCES

Achen, Christopher. 2006. "Evaluating Political Decision-Making Models." In Thomson, et al. 2006.

Allas, Tera, and Nikos Georgiades. 2001. "New Tools for Negotiators." *The McKinsey Quarterly* 2.

Altfeld, Michael F. 1983. "Arms Races?—And Escalation? A Comment on Wallace." *International Studies Quarterly* 27: 225–31.

Beck, Douglas, and Bruce Bueno de Mesquita. 1985. "Forecasting Political Decisions." In *Corporate Crisis Management*, ed. S. Andriole. Princeton, N.J.: Petrocelli Books.

Budge, Ian and Dennis J. Farlie. 1981. "Predicting Regime Change: A Cross-National Investigation with Aggregate Data 1950–1980." *Quality and Quantity* 15: 335–64.

Bueno de Mesquita, Bruce. 1990. "Multilateral Negotiations: A Spatial Analysis of the Arab-Israeli Dispute." *International Organization* (Summer): 317–40.

Bueno de Mesquita, Bruce. 2009. *The Predictioneer's Game*. New York: Random House.

Bueno de Mesquita, Bruce. 2011. "A New Model for Predicting Policy Choices: Preliminary Tests." *Conflict Management and Peace Science* forthcoming.

Bueno de Mesquita, Bruce, Feryal Cherif, George W. Downs, and Alastair Smith. 2005. "Thinking Inside the Box: A Closer Look at Democracy and Human Rights." *International Studies Quarterly*. 49(3): 439–57.

Bueno de Mesquita, Bruce, and Grace Iusi-Scarborough. 1988. "Forecasting the Nature of Political Settlement in Nicaragua." Paper presented at the Conference on Nicaragua: Prospects for a Democratic Outcome, sponsored by the Orkand Corporation, Washington, D.C., October.

Bueno de Mesquita, Bruce, and Chae-Han Kim. 1991. "Prospects for a New Regional Order in Northeast Asia." *Korean Journal of Defense Analysis* (Winter): 65–82.

Bueno de Mesquita, Bruce, Rose McDermott, and Emily Cope. 2001. "The Expected Prospects for Peace in Northern Ireland." *International Interactions* 27(2): 129–67.

Bueno de Mesquita, Bruce, David Newman, and Alvin Rabushka. 1985. *Forecasting Political Events*. New Haven: Yale University Press.

Bueno de Mesquita, Bruce, and A.F.K Organski. [1990] 1992. "A Mark in Time Saves Nein." *International Political Science Review* 13: 81–100.

Bueno de Mesquita, Bruce, and Alastair Smith. 2009. "Political Survival and Endogenous Institutional Change." *Comparative Political Studies* 42(2) (February): 167–97.

Bueno de Mesquita, Bruce, Alastair Smith, Randolph M. Siverson, and James D. Morrow. 2003. *The Logic of Political Survival*. Cambridge, Mass.: MIT Press.

Bueno de Mesquita, Bruce, and Frans Stokman, eds. 1994. *European Community Decision Making: Models, Applications, and Comparisons*. New Haven: Yale University Press.

Davenport, Christian. 2007. "State Repression and Political Order." *Annual Review of Political Science* 10: 1–23.

Dixon, Hugo, and Alexander Nicoll. 1999. "How Project Super Bowl Won the Day." *The Financial Times*, 23–24 January: 2.

Feder, Stanley. 1995. "Factions and Policon: New Ways to Analyze Politics." In *Inside the CIA's Private World: Declassified Articles from the Agency's Internal Journal,* 1955–1992, ed. H. Bradford Westerfield. New Haven: Yale University Press.

Feder, Stanley. 2002. "Forecasting for Policy Making on the Post-Cold War Period." *Annual Review of Political Science* 5: 111–25.

Green, Kesten. 2002. "Embroiled in a Conflict: Who Do You Call? *International Journal of Forecasting* 18(3): 389–95.

James, Patrick. 1998. "Rational Choice? Crisis Bargaining Over the Meech Lake Accord." Paper presented at the Annual Meeting of the Canadian Political Science Association, Charlottetown, Prince Edward Island, June." *Conflict Management and Peace Science* 16(2): 51–86.

Jervis, Robert. 1978. "Cooperation Under the Security Dilemma." *World Politics* 30: 167–214.

Koubi, Vally. 1994. "Military Buildups and Arms Control Agreements." *International Studies Quarterly* 38: 605–20.

Kugler, Jacek. 1987. "The Politics of Foreign Debt in Latin America: A Study of the Debtors' Cartel." *International Interactions* 13: 115–44.

Kugler, Jacek, Lewis W. Snider, and William Longwell. 1994. "From Desert Shield to Desert Storm: Success, Strife, or Quagmire?" *Conflict Management and Peace Science* 13: 113–48.

Morrow, James D., Bruce Bueno de Mesquita, and Samuel Wu. 1993. "Forecasting the Risks of Nuclear Proliferation: Taiwan as an Illustration of Method." *Security Studies* 2: 311–31.

McDermott, Rose, and Jacek Kugler. 2001. "Comparing Rational Choice and Prospect Theory: The U.S. Decision to Launch Operation Desert Storm in January 1991." *Journal of Strategic Studies* 24(3): 49–85.

Morgenthau, Hans. 1973. *Politics Among Nations: The Struggle for Power and Peace*, 5th ed. New York: Knopf.

Mousavi, Shabnam, and Hersh Shelfrin. Forthcoming. "Prediction Tools: Psychology, Politics, and Financial Market Regulation." *Journal of Risk Management in Financial Institutions*.

Newman, David, and Brian Bridges. 1994. "North Korean Nuclear Weapons Policy: An Expected Utility Study." *Pacific Focus* 9 (Fall): 61–80.

Organski, A.F.K., and Bruce Bueno de Mesquita. 1993. "Forecasting the 1992 French Referendum." In *New Diplomacy in the Post-Cold War World*, ed. Roger Morgan, J. Lorentzen and A. Leander. New York: St. Martin's Press.

Poe, Steven C., and C. Neal Tate. 1994. "Repression of Human Rights to Personal Integrity in the 1980s: A Global Analysis." *American Political Science Review* 88(4): 853–72.

Poe, Steven C., C. Neal Tate, and Linda Camp Keith. 1999. "Repression of the Human Right to Personal Integrity Revisited: A Global Cross-National Study

Covering the Years 1976–1993." *International Studies Quarterly* 43(2): 291–313.

Posner, Richard. 2002. *Public Intellectuals: A Study of Decline*. Cambridge, Mass.: Harvard University Press.

Ray, James, and Bruce M. Russett. 1996. "The Future as Arbiter of Theoretical Controversies: Predictions, Explanations, and the end of the Cold War." *British Journal of Political Science* 26: 441–70.

Richardson, Lewis Fry. 1960. *Statistics of Deadly Quarrels*. Pacific Grove, Calif.: Boxwood Press.

Richardson, Lewis Fry. 1978. *Arms and Insecurity: A Mathematical Study of the Causes and Origins of War*. Pacific Grove, Calif.: Boxwood Press.

Rotberg, Robert, and Theodore Rabb, eds. 1988. *The Origin and Prevention of Major Wars*. Cambridge: Cambridge University Press.

Schneider, Gerald, Daniel Finke, and Stefanie Bailer. 2010. "Bargaining Power in the European Union: An Evaluation of Competing Game-Theoretic Models." *Political Studies* 58(1): 85–103.

Schneider, Gerald, Nils Petter Gleditsch, and Sabine Carey. 2011. "Forecasting in International Relations: One Quest, Three Approaches." *Conflict Management and Peace Science* forthcoming.

Tetlock, Philip. 2005. *Expert Political Judgment*. Princeton: Princeton University Press.

Tetlock, Philip. 2009. "Reading Tarot on K Street." *The National Interest* (September/October): 57–67.

Thomson, Robert, Frans N. Stokman, Christopher H. Achen, and Thomas König, eds. 2006. *The European Union Decides*. Cambridge: Cambridge University Press.

Waltz, Kenneth. 1979. *Theory of International Politics*. Boston: McGraw Hill.

Wu, Samuel, and Bruce Bueno de Mesquita. 1994. "Assessing the Dispute in the South China Sea: A Model of China's Security Decision Making." *International Studies Quarterly* 38: 379–403.