December 2004

Value of expertise for forecasting decisions in conflicts

Kesten C. Green
Monash University

J. Scott Armstrong
University of Pennsylvania, armstrong@wharton.upenn.edu

Value of expertise for forecasting decisions in conflicts

Abstract
In important conflicts, people typically rely on experts' judgments to predict the decisions that adversaries will make. We compared the accuracy of 106 expert and 169 novice forecasts for eight real conflicts. The forecasts of experts using unaided judgment were little better than those of novices, and neither were much better than simply guessing. The forecasts of experts with more experience were no more accurate than those with less. Speculating that consideration of the relative frequency of decisions might improve accuracy, we obtained 89 forecasts from novices instructed to assume there were 100 similar situations and to ascribe frequencies to decisions. Their forecasts were no more accurate than 96 forecasts from novices asked to pick the most likely decision. We conclude that expert judgment should not be used for predicting decisions that people will make in conflicts. Their use might lead decision makers to overlook other, more useful, approaches.

Keywords
bad faith, framing, hindsight bias, methods, politics

Value of expertise for forecasting decisions in conflicts

Kesten C. Green,* Department of Econometrics and Business Statistics
Monash University, VIC 3800, Australia
E-mail: kesten@kestencgreen.com
Phone +61 3 990 55438
Fax +61 3 990 55474

J. Scott Armstrong, The Wharton School, University of Pennsylvania
Philadelphia, PA 19104
E-mail: armstrong@wharton.upenn.edu
Phone 610-622-6480
Fax 215-898-2534

December 7, 2004

Abstract

In important conflicts, people typically rely on experts' judgments to predict the decisions that adversaries will make. We compared the accuracy of 106 expert and 169 novice forecasts for eight real conflicts. The forecasts of experts using unaided judgment were little better than those of novices, and neither were much better than simply guessing. The forecasts of experts with more experience were no more accurate than those with less. Speculating that consideration of the relative frequency of decisions might improve accuracy, we obtained 89 forecasts from novices instructed to assume there were 100 similar situations and to ascribe frequencies to decisions. Their forecasts were no more accurate than 96 forecasts from novices asked to pick the most likely decision. We conclude that expert judgment should not be used for predicting decisions that people will make in conflicts. Their use might lead decision makers to overlook other, more useful, approaches.

Keywords: bad faith, framing, hindsight bias, methods, politics.

*Corresponding author.

Asking an expert to predict what will happen in a conflict seems a reasonable thing to do. The media find people to tell us what will happen when discussing conflicts such as the war on terrorism. In business, the CEO might ask the marketing manager to predict how competitors will respond to a new product launch or ask the human resources manager whether the offer of a 2% wage increase will deter a threatened strike. In the military, a general might ask his intelligence officer whether the enemy is likely to defend an outpost.

Evidence from surveys suggests that forecasts of decisions in conflicts are typically based on experts' unaided judgments (Armstrong, Brodie, and McIntyre 1987). Informal evidence that this is so abounds in everyday life and in the news. Winston Churchill observed that a politician should have "The ability to foretell what is going to happen… And to have the ability afterwards to explain why it didn't happen" (Adler 1965, p. 4). The same observation might be made of executives in business, the public sector, and the armed services.

While it is attractive to think that if we can find the right expert we can know what will happen, Armstrong (1980) in a review of evidence from diverse subject areas was unable to find evidence that expertise, beyond a modest level, improves experts' forecasting accuracy.

Some beliefs about the value of expertise

What do people think about the value of expertise when forecasting decisions in conflict situations? Prior to giving some talks on forecasting, we asked attendees for their opinions on the likely accuracy of experts' and novices' (university students') forecasts of decisions in conflicts. We told respondents that, for the purpose of our survey, they should assume the average accuracy of a random selection (chance) was 28%. Our talks were to academics and students at Lancaster University (19 usable responses), Manchester Business School (18), Melbourne Business School (6), Royal New Zealand Police College educators (4), Harvard Business School alumni (8), conflict management practitioners in New Zealand (7), and attendees at the International Conference on Organizational Foresight in Glasgow (15). A copy of the questionnaire we used is available at www.conflictforecasting.com.. We excluded all responses from those who expected accuracy to be less than 28% for any method as it seems implausible that the forecasts of any method would be worse than chance.

Our practitioners, forecasting experts, and miscellaneous academics had little faith in the predictions of novices, expecting their predictions to be accurate only 30% of the time – little different from chance. While they had greater confidence in experts, they expected only 45% of their forecasts to be accurate. Given that people often judge others' decisions in conflicts to be the result of stupidity or duplicity, we were surprised that our respondents expected that experts would typically be wrong when they make such predictions.

We suggest that prediction is difficult because conflicts tend to be too complex for people to think through in ways that realistically represent the situation. Parties in conflict often act and react many times, and change as a result of their interactions. There may be interactions within each party and there may well be more than two parties involved.

Tversky and Kahneman (1982b) suggested that when people are faced with complex situations, they are likely to resort to the heuristic of availability in order to judge the likelihood of outcomes. That is, they test their memories and judge an outcome likely when a similar outcome is easily recalled or imagined. For example, people nowadays tend to think it likely that war will end badly because the unceremonious withdrawal of US and allied troops from Vietnam is so easily recalled. There is, however, ample reason to be skeptical about whether the availability heuristic will lead to accurate predictions. For example, salient outcomes and the situations that gave rise to them are unlikely to be representative; quite the opposite. Unstructured reviews of the past are likely to offer poor guidance for the future (Fischhoff 1982, Harvey 2001).

Information processing is problematic. If we take Bayes's theorem as the standard, people tend to adjust their predictions less than they should when they receive new information (Edwards 1982). When they consider the likelihood of an outcome from a multistage process (Hitler invades Belgium, he succeeds, Britain declares war, Hitler attacks Britain) people have the opposite tendency: they act as though their best guesses of what will happen at early stages are certainties (Gettys, Kelly, and Peterson 1982).

Stewart (2001) found that judgmental forecasts are likely to be unreliable when (1) the task is complex, (2) there is uncertainty about the environment, (3) information acquisition is subjective, or (4) information processing is subjective. Stewart's four conditions for unreliability are likely to be met with the type of problems we are considering.

It is difficult for people to become better at predicting decisions in conflicts using unaided judgment because basic conditions for learning are typically absent. Timely and unambiguous feedback is uncommon, and opportunities for practise are rare (Arkes 2001). Feedback may be in the form of deliberately misleading information leaked by an adversary or the unreliable accounts of witnesses. Accurate feedback may be misinterpreted because experts misunderstand the situation (Einhorn 1982). Decision-makers may take action aimed at avoiding a predicted outcome thereby confounding feedback. Conflicts often occur over long periods of time and experts may have left before there is an outcome. Many experts will be faced with important conflicts only rarely and, in any case, conflicts are typically diverse and each one may appear more-or-less novel. Spurious correlations that support experts' theories can be readily constructed (Chapman and Chapman 1982, Jennings, Amabile, and Ross 1982).

Finally, Tetlock (1999) found that experts have excellent defenses against evidence that their forecasts are wrong, so that even in situations where conditions for learning are better than we have portrayed, experts may still fail to learn.

Robert McNamara (Morris 2003), Secretary of Defense under Presidents Kennedy and Johnson, refers to the "fog of war" in relation to conflicts he was involved in. We suggest that this term, which appears to have originated in the writings of Carl von Clausewitz (1993), might reasonably be applied to most conflict situations where unaided judgment is applied.

Research method

We recruited domain experts, conflict experts, and forecasting experts to predict the decisions made in eight diverse conflicts. The conflicts were real situations that were either obscure or were disguised in order to make recognition unlikely. The conflicts involved nurses striking for pay parity, football players wanting a bigger share of revenues, an employee resisting the down-grading of her job, artists demanding public financial support, a novel distribution arrangement with retailers, a hostile takeover attempt, a controversial investment proposal, and nations preparing for war. Each involved two or more interacting parties. The materials used in our research are available on www.conflictforecasting.com.

We allocated the conflicts to participants on the basis of their expertise. For example, we sent conflicts between employers and employees to industrial relations specialists, and we sent all eight conflicts to conflict management experts. Contact with participants was via email messages, and hence we had no control over the time they spent on the task, or whether they referred to other materials or other people.

We recruited novices to make predictions for the same situations (Green 2004). Materials were the same as for the experts but, instead of receiving the material by email, the students were paid to sit in lecture theatres and make their predictions. No attempts were made to match the backgrounds of the students with the subject matter of the conflicts and, unlike the experts who had discretion over which if any of the conflicts they made predictions for, the students were paid $20 only when they had provided forecasts for all their allocated conflicts.

Obtaining the forecasts

For each conflict, we provided participants with a set of between three and six decision options. We gave no instructions to participants on how they should make their predictions.

The way in which a problem is posed often affects judgmental predictions. One important distinction is whether a problem is framed as specific instance or a class of situations. For example, one might ask “how probable is it that the US will sign the Kyoto Protocol?” Alternatively, one could frame the problem as “in what proportion of cases would the US sign a treaty that would cause certain harm to the nation’s interests in return for uncertain benefits?” Kahneman and Tversky (1982a, 1982b) proposed that whereas people tend to think of situations as being “singular” (Kyoto Protocol) when they assess the likelihood of outcomes, their predictions would be more accurate if they used a “distributional” approach (international treaties) to assess likelihood. Tversky and Koehler (1994) postulated that the greater accuracy is a result of peoples’ tendency to consider alternatives in more detail. They suggest that people are prompted to think more about different ways that an outcome might occur when a problem is framed as a class of similar situations than when it is framed as a singular instance.

We conducted an experiment to compare the accuracy of unaided-judgment forecasts collected using a singular format with those collected by asking for frequencies of different decisions

across a set of hypothetical similar situations. We hypothesized that participants who were asked for frequencies might provide forecasts that were more accurate than those who were not.

Fifty-two participants, all university students, were paid the equivalent of US$20 to take part. We allocated them randomly between the singular and frequencies treatments. Each singular-treatment participant received a different sequence of four of the eight conflicts we used in our research and the frequencies-treatment sequences matched these. For each conflict, participants were given approximately 30 minutes to read the material and answer the questions before we collected their responses.

Four participants each claimed to recognize a situation, and we excluded their responses. Aside from the following forecasting questions, the treatments were identical.

Singular treatment question:

How was the stand-off between Localville and Expander resolved? (check one ✓ or %)

a. Expander's takeover bid failed completely [___]
b. Expander purchased Localville's mobile operation only [___]
c. Expander's takeover succeeded at, or close to, their August 14 offer price of $43-per-share [___]
d. Expander's takeover succeeded at a substantial premium over the August 14 offer price [___]

Frequencies treatment question:

Assume there are 100 situations similar to the one described, in how many of these situations would…

a. The takeover bid fail completely? [___] out of 100
b. The mobile operation alone be purchased? [___] out of 100
c. The takeover succeed at, or close to, the offer price? [___] out of 100
d. The takeover succeed at a substantial premium over the offer price? [___] out of 100

Findings

Expert versus novice judgment

Recall that our survey respondents expected experts' unaided-judgment forecasts to be substantially more accurate (45%) than those of novices (30%). This did not prove to be the case. The unaided experts' accuracy averaged only 32% across the conflicts used in our studies, little better than the average accuracy of 29% for novices' forecasts (Table 1). These results are consistent with evidence summarized in Armstrong (1985, pp. 91-96). There was little relationship between expertise and forecast accuracy. Neither group did appreciably better than chance. We used the permutation test for paired replicates (Siegel and Castellan 1988) to test the significance of the differences in accuracy between experts and chance. As a casual inspection of the data in Table 1 suggests, the differences across the eight conflicts are likely to have arisen by chance ($P = 0.30$, one-tail test). The test is 100% power-efficient as all the information is used (p. 100).

Table 1
Accuracy of unaided-judgment forecasts
Percent correct forecasts (number of forecasts)

	Chance	**By novices**		**By experts**	
Artists Protest	17	5	(39)	**10**	(20)
Distribution Channel	33	5	(42)	**38**	(17)
Telco Takeover	25	**10**	(10)	0	(8)
55% Pay Plan	25	**27**	(15)	18	(11)
Zenith Investment	33	29	(21)	**36**	(14)
Personal Grievance	25	**44**	(9)	31	(13)
Water Dispute	33	45	(11)	**50**	(8)
Nurses Dispute	33	68	(22)	**73**	(15)
Averages (unweighted)	28	29	(169)	32	(106)

Expert experience and accuracy

Is it possible to identify experts who are more likely than others to make accurate judgmental forecasts? One obvious way to assess this is to compare the accuracy of forecasts from experts with more experience with those from experts with less.

We asked expert participants to record the number of years experience they had as "a conflict management specialist". As a check, we also asked some of our novice participants and the responses were not surprising. Ninety-four percent of the university student participants who answered the question gave their experience as zero years, the rest claimed one or two years of such experience.

Common-sense expectations did not prove to be correct. The 57 forecasts of experts with less than five years experience were more accurate (36%) than the 48 forecasts of experts with more experience (29%).

Frequency responses and accuracy

We anticipated participants would be more accurate when asked to estimate the frequencies of outcomes for many similar situations. To our surprise, our university student participants who judged relative frequencies were no more likely to identify the actual decision than were participants who simply chose the decision they thought most likely. Both groups were correct for 33% of forecasts, averaged across conflicts (Table 2). The two sets of responses appear to follow the same pattern when looking across the situations (Spearman rank order correlation coefficient 0.59, $P < 0.10$; Siegel and Castellan 1988).

Table 2
Accuracy of novices' frequency and singular forecasts
Percent correct forecasts (number of forecasts)

	Chance	**Frequencies**		**Singular**		**Total**	
55% Pay Plan	25	0	(12)	9	(11)	4	(23)
Artists Protest	17	10	(10)	0	(11)	5	(21)
Distribution Channel	33	23	(13)	38	(13)	31	(26)
Personal Grievance	25	11	(9)	46	(13)	32	(22)
Telco Takeover	25	50	(12)	25	(12)	38	(24)
Zenith Investment	33	40	(10)	42	(12)	41	(22)
Water Dispute	33	67	(12)	42	(12)	54	(24)
Nurses Dispute	33	64	(11)	58	(12)	61	(23)
Averages (unweighted)	28	33	(89)	33	(96)	33	(185)

Of the 89 frequencies predictions, 54% summed to the specified total of 100; 35% totaled more than 100 and 11% less. It is arguable that, despite our intentions, the decision options we provided were not entirely mutually-exclusive or exhaustive and hence the failure of some participants' responses to add to 100 is not necessarily a failure of logic on their part. On the other hand, researchers have found that even with mutually exclusive and exhaustive lists of events, responses do not consistently sum to 1.0 or 100%, as people commonly fail to interpret probability or frequency scales in ways that researchers intend (Windschitl 2002).

Nonetheless, it seems reasonable to assume that our participants, who mostly had only three or four decision options to assess, allocated frequencies that were at least consistent with their ranking of the options' likelihoods. For our analysis, therefore, we used the decision with the highest frequency or probability, or the single decision chosen, as the forecast. We dropped ten observations where there was a tie.

When we excluded from our analysis responses that did not sum to 1.0 or 100, it made no difference to our conclusion – namely, that asking participants for frequencies did not improve accuracy. Across the conflicts, the average accuracy for frequencies responses was 29% (48 forecasts) compared to 32% (93) for singular treatment responses.

Discussion and conclusions

The various people we surveyed expected that it is difficult to forecast decisions in conflicts. Our evidence has shown that this is indeed the case. Most respondents nonetheless expected experts to be better forecasters than novices. They were wrong. Expertise did not improve accuracy. Neither experts nor novices did substantially better than guessing.

Our concerns that the wording of our forecasting tasks might have harmed accuracy proved unfounded. An analysis using only responses that conformed to the norms of probability theory led to the same conclusion.

The common use of judgment to forecast decisions in conflicts might lull decision makers into believing nothing further can be done. This would be unfortunate. There are no good grounds for decision makers to rely on experts' unaided judgements for forecasting decisions in conflicts. Misplaced reliance on expert judgment discourages experts and decision makers from investigating alternative approaches (Arkes 2001). While it is difficult to forecast in conflict situations, we have shown in Green (2004) and Green and Armstrong (2004) that it *is* possible to obtain substantially better forecasts.

Given the methods currently used in forecasting, to accuse expert advisors and political leaders of bad faith when their predictions about conflicts prove wrong does not seem justified. At best, such accusations are a product of hindsight bias. Inaccurate predictions are to be expected when experts use unaided judgment to predict how people will behave in conflicts.

Acknowledgements

We are grateful to Don Esslemont for helpful suggestions on various drafts of this article, and to Stuart Halpern for copy-editing. The article was also improved in response to probing questions from delegates at the 2003 and 2004 International Symposia on Forecasting and at the Institute of Mathematics and Its Applications' Conference on Conflict and Its Resolution, and from attendees at talks at RAND Organization, Warwick Business School, University College London, Monash University, and Melbourne Business School to whom we presented elements of the work reported here.

References

Adler, B. (Ed.) (1965). *The Churchill Wit*. New York: Coward-McCann.

Arkes, H. R. (2001). Overconfidence in judgmental forecasting, in Armstrong, J. S. (ed.), *Principles of forecasting*. Boston, MA: Kluwer Academic Publishers.

Armstrong, J. S., Brodie, R. J., & McIntyre, S. H. (1987). Forecasting methods for marketing: Review of empirical research. *International Journal of Forecasting*, **3**, 335-376.

Armstrong, J. S. (1980). The seer-sucker theory: The value of experts in forecasting. *Technology Review*, 83 (June/July), 18-24. Full text available at forecastingprinciples.com.

Armstrong, J. S. (1985). *Long-range forecasting*. New York: John Wiley. Full text available at forecastingprinciples.com.

Chapman, L. J., & Chapman, J. (1982). Test results are what you think they are, in Kahneman, D., Slovic, P. & Tversky, A. (eds.), *Judgment under uncertainty: heuristics and biases*. Cambridge, UK: Cambridge University Press.

Clausewitz, Carl von (1993). *On war*. Howard, M. & Paret, P. (ed./trans). New York: Alfred A. Knopf, "Everyman's Library" edition.

Edwards, W. (1982). Conservatism in human information processing, in Kahneman, D., Slovic, P. & Tversky, A. (eds.), *Judgment under uncertainty: heuristics and biases*. Cambridge, UK: Cambridge University Press.

Einhorn, H. J. (1982). Learning from experience and suboptimal rules in decision making, in Kahneman, D., Slovic, P. & Tversky, A. (eds.), *Judgment under uncertainty: heuristics and biases*. Cambridge, UK: Cambridge University Press.

Fischhoff, B. (1982). For those condemned to study the past: heuristics and biases in hindsight, in Kahneman, D., Slovic, P. & Tversky, A. (eds.), *Judgment under uncertainty: heuristics and biases*. Cambridge, UK: Cambridge University Press.

Gettys, C. F., Kelly, C., & Peterson, C. R. (1982). The best-guess hypothesis in multistage inference, in Kahneman, D., Slovic, P. & Tversky, A. (eds.), *Judgment under uncertainty: heuristics and biases*. Cambridge, UK: Cambridge University Press.

Green, K. C. (2004). Further evidence on game theory, simulated interaction, and unaided judgement for forecasting decisions in conflicts. Monash University Econometrics and Business Statistics Working Paper 18/04. Full text available at conflictforecasting.com

Green, K. C. & Armstrong, J. S. (2004). Structured analogies for forecasting. Monash University Econometrics and Business Statistics Working Paper 17/04. Full text available at conflictforecasting.com

Harvey, N. (2001). Improving judgment in forecasting, in Armstrong, J. S. (ed.), *Principles of forecasting*. Boston, MA: Kluwer Academic Publishers.

Jennings, D. L., Amabile, T. M., & Ross, L. (1982). Informal covariance assessment: data-based versus theory-based judgments, in Kahneman, D., Slovic, P. & Tversky, A. (eds.), *Judgment under uncertainty: heuristics and biases*. Cambridge, UK: Cambridge University Press.

Kahneman, D. & Tversky, A. (1982a). Intuitive prediction: biases and corrective procedures, in Kahneman, D., Slovic, P. & Tversky, A. (eds.), *Judgment under uncertainty: heuristics and biases*. Cambridge, UK: Cambridge University Press.

Kahneman, D. & Tversky, A. (1982b). Variants of uncertainty, in Kahneman, D., Slovic, P. & Tversky, A. (eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Morris, E. (2003). The fog of war: eleven lessons from the life of Robert S. McNamara. USA: Sony Pictures Classics. (Clips available for viewing at http://www.sonyclassics.com/fogofwar/)

Siegel, S. & Castellan, N. J. Jr. (1988). *Non-parametric Statistics for the Behavioral Sciences*, 2nd ed. Singapore: McGraw-Hill.

Stewart, T. R. (2001). Improving reliability in judgmental forecasts, in Armstrong, J. S. (ed.), *Principles of forecasting*. Boston, MA: Kluwer Academic Publishers.

Tetlock, P. E. (1999). Theory driven reasoning about possible pasts and probable futures: are we prisoners of our perceptions? *American Journal of Political Science*, **43**, 335-366.

Tversky, A. & Kahneman, D. (1982a). Judgments of and by representativeness, in Kahneman, D., Slovic, P. & Tversky, A. (eds.), *Judgment under uncertainty: heuristics and biases*. Cambridge, UK: Cambridge University Press.

Tversky, A. & Kahneman, D. (1982b). Availability: a heuristic for judging frequency and probability, in Kahneman, D., Slovic, P. & Tversky, A. (eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Tversky, A. & Koehler, D. J. (1994). Support theory: a nonextentional representation of subjective probability. *Psychological Review*, 101, 547-567.

Windschitl, P. D. (2002). Judging the accuracy of a likelihood judgement: the case of smoking risk. *Journal of Behavioral Decision Making*, 15, 19-35.