University of Pennsylvania
ScholarlyCommons

Marketing Papers

5-21-2007

The Ombudsman: Value of Expertise for Forecasting Decisions in Conflicts

Kesten C. Green
Monash University

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

Postprint version. Published in *Interfaces*, Volume 37, Issue 3, May 2007, pages 287-299.
Publisher URL: http://dx.doi.org/10.1287/inte.1060.0262

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/marketing_papers/48
For more information, please contact repository@pobox.upenn.edu.

The Ombudsman: Value of Expertise for Forecasting Decisions in Conflicts

Abstract
In important conflicts such as wars and labor-management disputes, people typically rely on the judgment of experts to predict the decisions that will be made. We compared the accuracy of 106 forecasts by experts and 169 forecasts by novices about eight real conflicts. The forecasts of experts who used their unaided judgment were little better than those of novices. Moreover, neither group's forecasts were much more accurate than simply guessing. The forecasts of experienced experts were no more accurate than the forecasts of those with less experience. The experts were nevertheless confident in the accuracy of their forecasts. Speculating that consideration of the relative frequency of decisions across similar conflicts might improve accuracy, we obtained 89 sets of frequencies from novices instructed to assume there were 100 similar situations. Forecasts based on the frequencies were no more accurate than 96 forecasts from novices asked to pick the single most likely decision. We conclude that expert judgment should not be used for predicting decisions that people will make in conflicts. When decision makers ask experts for their opinions, they are likely to overlook other, more useful, approaches.

applications, bargaining, behavior, competitive strategy, decision making, decision analysis, defense, effectiveness/performance, forecasting, foreign policy, leadership, military, organizational studies, strategy, tactics

Comments
Postprint version. Published in *Interfaces*, Volume 37, Issue 3, May 2007, pages 287-299.
Publisher URL: http://dx.doi.org/10.1287/inte.1060.0262

This journal article is available at ScholarlyCommons: http://repository.upenn.edu/marketing_papers/48

Interfaces
Vol. 00, No. 0, Xxxxx-Xxxxx 2007, pp. 1-12
issn 0092-2102 _ eissn 1526-551X _ 07 _ 0000 _ 0001

informs®

doi 10.1287/inte.1060.0262
© 2007 INFORMS

The Ombudsman: Value of Expertise for Forecasting Decisions in Conflicts

Kesten C. Green
Department of Econometrics and Business Statistics, Monash University, Victoria 3800, Australia,
kesten@kestencgreen.com

J. Scott Armstrong
The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104,
armstrong@wharton.upenn.edu

In important conflicts such as wars and labor-management disputes, people typically rely on the judgment of experts to predict the decisions that will be made. We compared the accuracy of 106 forecasts by experts and 169 forecasts by novices about eight real conflicts. The forecasts of experts who used their unaided judgment were little better than those of novices. Moreover, neither group's forecasts were much more accurate than simply guessing. The forecasts of experienced experts were no more accurate than the forecasts of those with less experience. The experts were nevertheless confident in the accuracy of their forecasts. Speculating that consideration of the relative frequency of decisions across similar conflicts might improve accuracy, we obtained 89 sets of frequencies from novices instructed to assume there were 100 similar situations. Forecasts based on the frequencies were no more accurate than 96 forecasts from novices asked to pick the single most likely decision. We conclude that expert judgment should not be used for predicting decisions that people will make in conflicts. When decision makers ask experts for their opinions, they are likely to overlook other, more useful, approaches.

Key words : applications; bargaining; behavior; competitive strategy; decision making; decision analysis; defense; effectiveness/performance; forecasting; foreign policy; leadership; military; organizational studies; strategy; tactics.

Asking an expert to predict what will happen in a conflict seems to be a reasonable thing to do. For example, the media find professors and politicians to tell us what will happen when discussing conflicts such as the war on terrorism. In business, a CEO might ask the company's marketing manager to predict competitor response to a new-product launch or ask the human resources manager whether offering a two-percent wage increase will deter a threatened strike. In the military, a general might ask an intelligence officer if 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 et al. 1987). Informal evidence that this is true abounds. 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 an expert's ability to forecast accurately.

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 talks about 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 that those asked to make predictions had been presented with

descriptions of several different conflicts and were asked to choose from between three and six possible decisions such that the expected accuracy from choosing randomly across the full set of conflicts was 28 percent. This percentage is the average chance of a correct prediction for the eight conflicts we used in our research, or _1/6 + 1/4 + 1/4 + 1/4 + 1/3 + 1/3 + 1/3 + 1/3_/8 ∗ 100. By asking respondents to adopt 28 percent chance as the value of chance when they made their assessments, we are able to make meaningful comparisons between our research findings and their accuracy expectations.

We conducted our surveys prior to giving talks 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 our questionnaire is available at www.conflictforecasting.com. We excluded 27 responses from people who expected accuracy to be less than 28 percent for any method because it seemed implausible to us that the forecasts of any method would, on average, be worse than chance. If a method really were worse than chance, the forecaster could eliminate the decision predicted by the method and choose another one at random, thereby obtaining forecasts that were more accurate than chance.

Our practitioners, forecasting experts, and miscellaneous academics had little faith in the judgment of novices, expecting their predictions to be accurate only 30 percent of the time—little better than chance. The respondents had greater confidence in experts—66 percent expected them to be more accurate than novices, whereas only 9 percent expected novices to be more accurate. Despite their greater faith in experts, respondents expected only 45 percent of experts' forecasts to be accurate. If the responses we excluded were included, the average expectations would be 30 percent for novices and 42 percent for experts, rather than 30 percent and 45 percent, respectively.

We suggest that accurate prediction is difficult because conflicts tend to be too complex for people to think through in ways that realistically represent their actual progress. Parties in conflict often act and react many times, and change because of their interactions. In addition, there may be interactions within each party, and there may be more than two parties involved.

Tversky and Kahneman (1982) suggested that when people are faced with complex situations, they are likely to resort to the heuristic of availability to judge the likelihood of outcomes. That is, they test their memories and judge an outcome likely when they can easily recall or imagine a similar one. For example, some people tend to think it likely that new wars will end badly because they have a vivid memory of the unceremonious withdrawal of US and allied troops from Vietnam (Kagan 2005). 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. Unstructured reviews of the past are likely to offer poor guidance for the future (Fischhoff 1982, Harvey 2001).

How people process information 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 (e.g., Hitler invades Belgium, he succeeds, Britain declares war, Hitler attacks Britain), people have the opposite tendency: they act as if their best guesses of what will happen at early stages are certainties (Gettys et al. 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. Problems of the type we are considering are likely to meet Stewart's four conditions for unreliability.

It is difficult for people to improve 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 practice are rare (Arkes 2001). Feedback may include misleading information that an adversary has disseminated or the unreliable accounts of witnesses. Accurate feedback may be misinterpreted because experts have misunderstood the

situation (Einhorn 1982). Decision makers may act to avoid a predicted outcome, thereby confounding feedback. Conflicts often occur over long periods of time, and those responsible for predicting an outcome may no longer be present when the actual outcome occurs. Many experts rarely face important conflicts. For those who do, each conflict may be unique. Experts can readily construct spurious correlations to support their theories (Chapman and Chapman 1982, Jennings et al. 1982).

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

Robert McNamara (Morris 2003), Secretary of Defense under Presidents Kennedy and Johnson, referred to the "fog of war " in relation to conflicts in which he was involved. We suggest that this term, which appears to have originated in the writings of Prussian Major General Carl von Clausewitz in 1832 (von Clausewitz 1993), might reasonably be applied to most conflict situations in which decision makers use their unaided judgment to make predictions.

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 for which accurate forecasts might reasonably have been expected to save money or lives. We disguised conflicts that were not obscure to make recognition of the real situation unlikely. We chose conflicts for their diversity and because we could get good information about them. The conflicts involved nurses striking for pay parity, football players seeking a bigger share of revenues, an employee resisting the downgrading of her job, artists demanding public financial support, a novel distribution arrangement that a manufacturer proposed to retailers, a hostile takeover attempt, a controversial investment proposal, and nations preparing for war. Each involved two or more interacting parties. The materials we used in our research are available on conflictforecasting.com.

We allocated the conflicts to expert participants based on 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. Because we used e-mail to contact participants, we had no control over how much time they spent on the task, or whether they referred to other materials or consulted other people.

We recruited novices to make predictions for the same situations (Green 2005) and provided them with the same materials. Rather than sending them the material by e-mail, we paid the students to make their predictions while they sat in lecture theatres. We did not attempt to match students' knowledge and experience with the subject matter of the conflicts. Unlike the experts who had discretion over the conflicts for which they made predictions, the students were paid only when they had provided forecasts for all of the conflicts that we had allocated to them.

Obtaining the Forecasts

For each conflict, we provided participants with a set of between three and six decision options. We gave them no instructions 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 a 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, b) proposed that, whereas people tend to think of situations as being "singular " when they assess the likelihood of outcomes (e.g., Kyoto Protocol signature), their predictions would be more accurate if they used a "distributional" approach (e.g., international treaty signatures) to assess likelihood. Kahneman and Lovallo (1993) used the term "outside view" when they presented evidence on the superiority of a distributional approach. Tversky and Koehler (1994) postulated that the greater accuracy is a result of peoples' tendency to consider alternatives in more detail. They suggested 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. Cosmides and Tooby (1996) found evidence for the proposition that people have innate mechanisms for storing and manipulating frequency information.

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.

We paid 52 university students the equivalent of US$20 to take part in the experiment and allocated them randomly between the singular and frequencies treatments. Each singular-treatment participant received a different sequence of four of the eight conflicts that we used in our research; we gave matching sequences to the frequencies-treatment participants. We allowed participants approximately 30 minutes to read the material and answer the questions for each conflict.

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

Singular treatment question:
How was the standoff 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 that 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 vs. Novice Judgment

Our survey respondents expected experts' unaided-judgment forecasts to be substantially more accurate (45 percent) than those of novices (30 percent). This expectation was not borne out. The unaided experts' forecast accuracy averaged only 32 percent across the conflicts used in our studies, little better than the average accuracy of 29 percent for novices' forecasts (Table 1). Neither group did appreciably better than chance. These results are consistent with evidence that Armstrong summarized (1985, pp. 91-96).

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 across the eight conflicts. As a casual inspection of the data in Table 1 suggests, the differences are quite likely to have arisen by chance (P = 0_30, one-tail test). The test is 100 percent power-efficient because it uses all the information (Siegel and Castellan 1988, p. 100).

Expert Experience and Accuracy

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

We asked expert participants to record their years of experience as "a conflict management specialist."

	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)	2829 (169)	32 (106)	

Table 1: We show the percentage accuracy of unaided judgment forecasts (numbers of forecasts are in parentheses).

As a check, we also asked some of our novice participants the same question. Their responses were as expected: 94 percent of the university-student participants who answered the question reported that they had no experience; 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 percent) than the 48 forecasts of experts with more experience (29 percent).

We also asked our expert participants to rate their experience with conflicts similar to the one they were examining using a scale from 0 to 10. Those who considered they had little experience with similar conflicts (they gave themselves ratings of 0 or 1) were as equally accurate at 34 percent (72 forecasts) as those who gave themselves higher ratings (32 forecasts).

Expert Confidence and Accuracy

We wondered whether experts' confidence in their individual forecasts could be used to identify accurate forecasts. On the other hand, their confidence might be misplaced when the forecasting problems are difficult. We asked our expert participants:

How likely is it that taking more time would change your forecast?

{0 = almost no chance _1/100_ _ _ _ 10 = practically certain _99/100_} 0-10.

While it is possible that the experts might have reasoned that they were unlikely to change a forecast given more time because they did not expect their forecast to be better than guessing, the fact of their participation and our evidence on accuracy expectations suggests that this was not the case. We interpret the experts' responses to this question as a measure of their confidence in the accuracy of their forecasts. We compared the accuracy of forecasts in which experts had high confidence with those in which they had less confidence. When experts assessed the likelihood that they would change their forecasts if given more time as between 0 and 2 out of 10, i.e., no more than 0.2 probability of change, we coded the forecasts as "high confidence." All other forecasts we coded as "low confidence." Using unweighted averages across the conflicts, the 68 high-confidence forecasts were less accurate (at 28 percent) than the 35 low-confidence forecasts (at 41 percent).

We also compared the confidence that the experts expressed in their forecasts that turned out to be accurate with their confidence in forecasts that turned out to be inaccurate. There were six conflicts for which we had both accurate and inaccurate forecasts and for which there were no half-accurate forecasts (the "distribution channel" conflict offered the option "c. Either a or b" and we coded the nine such responses as 0.5). Using unweighted averages across the six conflicts, we found that the experts assessed the probability that they would change the 27 accurate forecasts as 0.25, and that they would change the 51 inaccurate forecasts as 0.17, again showing a lack of relationship between confidence and accuracy.

Frequency Responses and Accuracy

We expected that forecasts would be more accurate when we asked our participants to estimate the frequencies of outcomes for many similar situations. Our university-student participants who judged relative frequencies were no better at identifying the actual decision than were those who simply chose the decision they thought most likely. Averaged across conflicts, 33 percent of forecasts from both the frequencies and singular treatments were accurate (Table 2). Further, the accuracy figures for the two groups 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).

Of the 89 frequencies predictions, 54 percent summed to the total of 100 that was specified in the frequencies-treatment question; 35 percent totaled more

	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)

Table 2: We show the percentage accuracy of novices' frequency and singular forecasts (numbers of forecasts are in parentheses).

than 100, and 11 percent less than 100. It is arguable that, despite our intentions, the decision options we provided were not entirely mutually exclusive or exhaustive, and 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 percent because 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 in most cases 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 10 observations in which there was a tie.

When we excluded responses that did not sum to 1.0 or 100, it did not change our conclusion that asking participants for frequencies did not improve accuracy. Across the conflicts, the average accuracy for frequencies responses was 29 percent (48 forecasts) compared with 32 percent (93 forecasts) for singular-treatment responses.

Discussion and Conclusions

The people we surveyed expected that forecasting decisions in conflicts would be difficult. Our findings confirmed this. 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 our instructions to participants might have harmed accuracy proved unfounded: asking for an assessment of the relative frequency of decisions across similar situations did not help. An analysis using only responses that conformed to the norms of probability theory led to the same conclusion. We suggest that the complexity of conflict situations means that people tend to view each one as more-or-less unique and, therefore, do not store or recall frequency information in the way that they do for simpler situations such as rainy days in April, or the presence of speed cameras on their routes to work.

There are no good grounds for decision makers to rely on experts' unaided judgments for forecasting decisions in conflicts. Such reliance discourages experts and decision makers from investigating alternative approaches (Arkes 2001).

While it is difficult to accurately forecast decisions in conflict situations, we have shown in Green (2005) and Green and Armstrong (2004) that it is possible to obtain substantially better forecasts. Green (2005) found that simulated interaction, a type of role playing for forecasting behavior in conflicts, reduced error by 47 percent when compared with game-theory experts' forecasts. (Role players were mostly undergraduate students.) In Green and Armstrong (2004), we asked experts to recall and analyze information on similar situations from the past using a method we called structured analogies. When experts were able to think of at least two analogies, forecast error was reduced by 39 percent compared to chance accuracy.

While expert advisors and political leaders use unaided judgment to forecast, it is unreasonable to accuse them of bad faith when their predictions about conflicts prove wrong. We should expect inaccurate predictions when experts use unaided judgment to forecast how people will behave in conflicts.

Acknowledgments

We are grateful to Paul Goodwin for organizing the special section for this article and to Robyn Dawes, Don Esslemont, Jonathan J. Koehler, and Lee Ross for their helpful suggestions on various drafts of this article. We are also grateful to Stuart Halpern, Bryan LaFrance, Alice Barrett Mack, and Alexandra Yordanova for their editing help. The article was 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 people at Rand Corporation, the CIA's Sherman Kent School, 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. Coward-McCann, New York.

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

Armstrong, J. S. 1980. The seer-sucker theory: The value of experts in forecasting. Tech. Rev. 83(June/July) 18-24. forecastingprinciples.com.

Armstrong, J. S. 1985. Long-Range Forecasting. John Wiley & Sons, New York. forecastingprinciples.com.

Armstrong, J. S., R. J. Brodie, S. H. McIntyre. 1987. Forecasting methods for marketing: Review of empirical research. Internat. J. Forecasting 3 335-376.

Chapman, L. J., J. Chapman. 1982. Test results are what you think they are. D. Kahneman, P. Slovic, A. Tversky, eds. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Cosmides, L., J. Tooby. 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition 58 1-73.

Edwards, W. 1982. Conservatism in human information processing. D. Kahneman, P. Slovic, A. Tversky, eds. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Einhorn, J. H. 1982. Learning from experience and suboptimal rules in decision making. D. Kahneman, P. Slovic, A. Tversky, eds. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Fischhoff, B. 1982. For those condemned to study the past: Heuristics and biases in hindsight. D. Kahneman, P. Slovic, A. Tversky, eds. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Gettys, C. F., C. Kelly, C. R. Peterson. 1982. The best-guess hypothesis in multistage inference. D. Kahneman, P. Slovic, A. Tversky, eds. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Green, K. C. 2005. Game theory, simulated interaction, and unaided judgement for forecasting decisions in conflicts: Further evidence. Internat. J. Forecasting 21 463-472. conflictforecasting.com

Green, K. C., J. S. Armstrong. 2004. Structured analogies for forecasting. Working Paper 17/04, Department of Econometrics and Business Statistics, Monash University, Victoria, Australia. conflictforecasting.com.

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

Jennings, D. L., T. M. Amabile, L. Ross. 1982. Informal covariation assessment: Data-based versus theory-based judgments. D. Kahneman, P. Slovic, A. Tversky, eds. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Kagan, F. W. 2005. Iraq is not Vietnam. Policy Rev. 134. policyreview.org.

Kahneman, D., D. Lovallo. 1993. Timid choices and bold forecasts: A cognitive perspective on risk taking. Management Sci. 39 17-31.

Kahneman, D., A. Tversky. 1982a. Intuitive prediction: Biases and corrective procedures. D. Kahneman, P. Slovic, A. Tversky, eds. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Kahneman, D., A. Tversky. 1982b. Variants of uncertainty. D. Kahneman, P. Slovic, A. Tversky, eds. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Morris, E. 2003. The fog of war: Eleven lessons from the life of Robert S. McNamara. USA: Sony Pictures Classics. Documentary film.

Siegel, S., N. J. Castellan, Jr. 1988. Nonparametric Statistics for the Behavioral Sciences, 2nd ed. McGraw-Hill, Singapore.

Stewart, T. R. 2001. Improving reliability in judgmental forecasts. J. S. Armstrong, ed. Principles of Forecasting. Kluwer Academic Publishers, Boston, MA.

Tetlock, P. E. 1999. Theory-driven reasoning about possible pasts and probable futures in world politics: Are we prisoners of our perceptions? Amer. J. Political Sci. 43(2) 335-366.

Tversky, A., D. Kahneman. 1982. Availability: A heuristic for judging frequency and probability. D. Kahneman, P. Slovic, A. Tversky, eds. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Tversky, A., D. J. Koehler. 1994. Support theory: A nonextensional representation of subjective probability. Psych. Rev. 101 (4) 547-567.

von Clausewitz, Carl. 1993. On War. M. Howard, P. Paret, ed./trans. Alfred A. Knopf, Everyman's Library ed., New York.

Windschitl, P. D. 2002. Judging the accuracy of a likelihood judgement: The case of smoking risk. J. Behavioral Decision Making 1519-35.

Comment: Combating Common Sense and Meeting Practitioner Needs

Shelley A. Kirkpatrick

Homeland Security Institute, 2900 South Quincy Street, Suite 800, Arlington, Virginia 22206, shelley.kirkpatrick@hsi.dhs.gov

Green and Armstrong discuss the accuracy of two forecasting methods—simulated interaction and structured analogies. This study, in conjunction with their previous research, provides compelling evidence that each of these methods yields more accurate forecasts than experts' unaided judgment (Green 2002, Green and Armstrong 2004).

I am a principal analyst at the Homeland Security Institute (HSI). The views I express in this article are my own and do not necessarily reflect HSI opinion or policy.

In my experience as a scientist practitioner who has worked with the intelligence, defense, and homeland security (IDHS) communities to assess the behavior of adversary groups and leaders, I have encountered many perspectives on expert judgment. To illustrate, I describe three views.

(1) Experts can address all problems: This is the view that we can accurately address national and homeland security issues, including conflict situations, by asking experts. Sometimes, we ask a group of experts to arrive at a group forecast. Other times, we seek a range of viewpoints, e.g., to identify new vulnerabilities to critical infrastructure. We seek expertise using many methods, including focus groups, panel discussions, conferences, meetings, and informal interactions.

(2) We cannot forecast all problems: Some problems are undefined or too complex, or we lack expertise about them. This is related to the view that, because experts are not always right, we should consult many experts. When one must consult an expert—such as when there is little objective data available and small changes in the environment—it is recommended that many, rather than few experts be asked to provide a judgment (Armstrong 1985). However, experts typically do not provide decision makers with quantitative forecasts, thus forcing decision makers to integrate a variety of qualitative viewpoints.

(3) We can model and forecast problems: Rather than relying on expert judgment, we can use modeling to quantify problems and yield a forecast. Quantitative modeling approaches often assume that the individuals we model operate with perfect rationality. For example, research has yielded no evidence that terrorists are mentally ill. However, rationality from a terrorist's perspective usually differs from rationality as perceived by a US citizen (Sageman 2004). Clearly, we cannot model all problems. Thomas Schelling, winner of a Nobel prize in economics for his work on game theory, states that game theory is less useful for analyzing how to deter terrorists from using nuclear weapons because "it is difficult to figure out what their objectives are" (Henderson 2005). Still, it is possible to apply current forecasting principles to the problem of terrorism (Green 2004). According to Heuer (1999), there is no failure to collect intelligence data, only failure to analyze it.

I present some ideas on the value of unaided expert judgment, and then follow with suggestions for future research to help practitioners develop forecasts.

Combating Common Sense

Examples abound of common-sense ideas that research does not support. The use of unstructured interviews to select new employees (Schmidt and Zimmerman 2004) is one illustration. In an unstructured interview, the interviewer asks different questions of different applicants or asks the same questions in a different order. In a structured interview, the interviewer asks all applicants the same questions, in the same order. The interviewer usually determines the questions, which are all intended to determine job-relevant abilities.

Research, including several meta-analyses, on unstructured interviews consistently finds it to be less accurate than structured interviews (Huffcutt and Arthur 1994, McDaniel et al. 1994, Wiesner and Cronshaw 1988) in predicting job performance. Despite these findings, interviewers commonly use the unstructured interview for several reasons. First, managers like unstructured interviews because they require little or no preparation. Second, managers frequently have already decided that the applicant is qualified but want to appraise qualities, such as communication skills, that are not always apparent on a resume. Third, applicants expect unstructured interviews and are familiar with an unstructured format.

Extrapolating from this example, we can find clues on why relying on expert judgment seems reasonable. I present these ideas to explain why the findings of Armstrong and Green appear counterintuitive, not to argue against their findings:

—Decision makers can engage experts in two-way conversation. Such dialogue enables experts to explain their forecasts and decision makers to improve their understanding of the problem.

—Experts can determine the decision maker's requirements, making future interactions with the decision maker more efficient.

—Experts are thought to arrive at forecasts, especially of new problems, quickly. Compared to an empirical study or analytical process, experts simply arrive at a judgment or decision; they do not go through an empirical process of designing a study,

collecting data, performing analyses, and drawing conclusions. Structured methods require time for data collection, analysis, interpretation of results, and communication of findings to the decision maker.

—The use of trusted experts enhances the forecast security and secrecy. Structured methods may require the involvement of more people, and results obtained from software-based forecasting methods can be copied and stolen. Even knowledge of forecast-data requirements can provide an adversary with information about the forecast.

—It is difficult to question or challenge a forecast without knowing how the expert arrived at it. Decision makers who prefer an expert approach or trust the judgment of a particular expert are often unlikely to ask the expert for an explanation of the judgment. Indeed, the expert may not be able to explain all of the factors considered in making the judgment.

It is difficult to convince people that their common-sense ideas are wrong. Rather than trying to do so, perhaps we should try to give decision makers a better understanding of the real value of unaided expert judgment. For example, such judgments may be useful in improving the decision maker 's understanding of a forecasting situation but not helpful when a specific forecast is required.

Meeting Practitioner Needs

I believe that Green and Armstrong are a positive example of researchers who strive to create new and useful knowledge for practitioners. Their websites (www.conflictforecasting.com and www.forecastingprinciples.com) are excellent resources for scientists and practitioners. I propose some ideas that they, and others, might consider for future research.

Their articles and websites provide descriptions of their methodologies and guidance for applying them. However, they may still leave practitioners uncertain on how to use the methodologies in their specific conflict situations. I encourage Green and Armstrong to continue to research the implementation of their methodologies, and thus to facilitate their practical use. Expanding their research to new problem sets and new study participants, for example, would be one way to demonstrate the broader applicability of their methods. Practitioners could serve as partners in the research process, such as by assisting in developing conflict situations that have external validity. I also suggest that they use their websites as a forum for practitioners and researchers to share role-playing instructions, new conflict scenarios, and lessons learned when applying the methodologies.

Finally, many subject-matter experts do not have training in developing a forecast in a structured manner. Therefore, I suggest a slightly different line of research to focus on training experts. In addition to determining ways to obtain accurate forecasts without using experts, finding ways to train subject-matter experts in forecasting may prove valuable.

References

Armstrong, J. S. 1985. Long-Range Forecasting: From Crystal Ball to Computer, 2nd ed. John Wiley & Sons, New York.

Green, K. C. 2002. Forecasting decisions in conflict situations: A comparison of game theory, role-playing, and unaided judgment. Internat. J. Forecasting 18 321-344.

Green, K. C. 2004. Better predictions can help defeat terrorism. Unpublished working paper, Monash University, Victoria, Australia.

Green, K. C., J. S. Armstrong. 2004. Structured analogies for forecasting. Working paper, Monash University, Victoria, Australia. Retrieved December 15, 2005 http://mktg-sun.wharton.upenn.edu/forecast/Conflicts/PDF%20files/Structured_Analogies107.pdf.

Henderson, N. 2005. Retired U-Md. economist wins Nobel Prize. Washington Post (October 11) A1.

Heuer, R. J., Jr. 1999. Psychology of Intelligence Analysis. Retrieved November 11, 2004 https://www.cia.gov/csi/books/19104/index.html.

Huffcutt, A. I., W. Arthur. 1994. Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. J. Appl. Psych. 79 184-190.

McDaniel, M. A., D. L. Whetzel, F. L. Schmidt, S. D. Maurer. 1994. The validity of employment interviews: A comprehensive review and meta-analysis. J. Appl. Psych. 79 599-616.

Sageman, M. 2004. Understanding Terror Networks. University of Pennsylvania Press, Philadelphia, PA.

Schmidt, F. L., R. D. Zimmerman. 2004. A counterintuitive hypothesis about employment interview validity and some supporting evidence. J. Appl. Psych. 89 553-561.

Wiesner, W. H., S. F. Cronshaw. 1988. The moderating impact of interview format and degree of structure on the validity of the employment interview. J. Occupational Psych. 61 275-290.

Interfaces 00(0), pp. 1-12, © 2007 INFORMS

Comment: Experts Who Don't Know They Don't Know

Jonathan J. Koehler
McCombs School of Business, University of Texas at Austin, 1 University Station B6500, Austin, Texas 78712-0212,
koehler@mail.utexas.edu

Sadly, the conclusion that Green and Armstrong reach—that experts should not be used for predicting the conflict outcomes—is not a surprise. Decades ago, Armstrong taught us that expertise beyond a minimal level does not improve judgmental accuracy across a variety of domains (Armstrong 1980). More recently, Tetlock (2005) drove home that point in a study of hundreds of political experts who made thousands of forecasts over many years. Like Green and Armstrong, Tetlock found the expert forecasts to be frequently inaccurate. In support of Armstrong's previous work, Tetlock suggests that avid readers of The New York Times should be able to predict political events as well as highly trained experts.

Green and Armstrong also demonstrate that nonprofessionals mistakenly expect superior performance from experts relative to what they expect from novices. Although it is true that neither novices nor experts were more accurate than guessing in eight conflict-prediction tasks, most study participants did not begin with high expectations of the experts. Participants expected experts to be accurate 45 percent of the time in tasks in which random guessing would yield a success rate of approximately 28 percent. Although these expectations were higher than chance, they are hardly a high endorsement for the perceived value of using expert forecasters.

However, if people really believe that experts are not good at predicting the future, why do we clamor for their views? Perhaps, we find it comforting to be with those who are knowledgeable about things that concern us. By speaking to our concerns, experts may justify our anxieties. Perhaps, experts help us to organize problems in our minds by laying out the advantages and disadvantages of the options we face. Or, when we ourselves must make decisions, perhaps experts function largely as convenient sources of blame for decisions that turn out badly (e.g., poor investment choices).

A question that may be more interesting than why we clamor for predictions from experts who disappoint is why experts continue to offer their faux expertise. The answer seems obvious: Experts predict because we ask them and reward them well for doing so. Fame, influence, and riches are the spoils of those who answer the media's incessant calls for forecasting expertise. However, I suspect that most experts genuinely believe in their forecasting skills. My suspicion may seem naïve in the face of consistent evidence that shows expert forecasters struggle to outperform novice forecasters and chance. Surely the experts know the data. They must know their own dismal records. Or, do they? My hunch is that they do not think their forecasting records are bad. Quite the contrary, they may believe that their records are outstanding.

Psychological research shows that people seek, recall, focus upon, and interpret evidence in ways that reinforce existing beliefs (Nisbett and Ross 1980). These cognitive biases reinforce our initial beliefs and prevent us from having to admit error or concede intellectual ground. If conflict experts believe that they are quite good at forecasting the resolution of certain types of conflicts, they may sustain their faith in their forecasting skills by remembering their correct calls and misremembering their failures. Or, perhaps, they interpret and encode failures as successes. World events are complicated, and deciding whether a political forecast (as opposed to a weather forecast or a sports-contest forecast) is or is not correct can be a matter of judgment or wish. Were the experts and politicians who said that former Iraqi leader Saddam Hussein possessed weapons of mass destruction immediately prior to the start of the 2003 United States-Iraq war correct? Most people think they were wrong. Others disagree, noting that Saddam Hussein did have those weapons at one time, that he used

them against his own people, and that he had the desire and means to obtain such weapons again. This defense is an example of what some philosophers refer to as a "fallacy of diversion" (Damer 1995), i.e., an attempt to maneuver oneself into a more advantageous or less embarrassing intellectual position by focusing on peripheral matters. This may insulate forecasters from having to contemplate, let alone concede, error.

Even when experts do concede forecast error, they may not alter their beliefs about their forecasting skills because they may find ways to minimize the import of their errors. As Tetlock (2005) documents in his study of political forecasters, experts find ways to avoid conceding error—even when faced with an outcome other than they predicted. Paraphrasing Tetlock's detailed discussion, common defenses of failed predictions include: (1) I was just off on timing—my predictions will eventually be borne out; (2) An improbable event occurred that changed the outcome; (3) My reasoning was accurate; and (4) My error was the lesser of the two errors that one could have made.

Green and Armstrong conclude on an optimistic note. They cite some of their other research, which shows that conflict-forecasting errors can be reduced when forecasters engage in role playing and draw upon analogies from previous conflicts. Until these and other decision aids are fully developed and in the cultural mainstream, we would be wise to bear in mind the two types of forecasters John Kenneth Galbraith identified: "Those who don't know, and those who don't know they don't know."

References

Armstrong, J. S. 1980. The seer-sucker theory: The value of experts in forecasting. Tech. Rev. 83 16-24.

Damer, T. E. 1995. Attacking Faulty Reasoning, 3rd ed. Wadsworth Publishing Co., Belmont, CA.

Nisbett, R., L. Ross. 1980. Human Inference: Strategies and Shortcomings of Social Judgment. Prentice-Hall, Englewood Cliffs, NJ.

Tetlock, P. E. 2005. Expert Political Judgment: How Good Is It? How Can We Know? Princeton University Press, Princeton, NJ.

Comment: Factors Promoting Forecasting Accuracy Among Experts: Some Multimethod Convergence

Philip E. Tetlock

Haas School of Business, University of California, Berkeley, 545 Student Services Building #1900, Berkeley, California, 94720-1900, tetlock@haas.berkeley.edu

The findings that Green and Armstrong report are compatible with many findings that I discussed in my recent book, Expert Political Judgment: How Good Is It? How Can We Know? (Tetlock 2005). Like Green and Armstrong, I found little support for the usual hypotheses about factors often believed to influence the accuracy of experts' predictions. When I examined approximately 28,000 predictions that 280 experts made on the political and economic futures of approximately 60 countries, I too found no difference in the accuracy of forecasts from: (1) experts versus dilettantes; (2) those with more experience and those with less; (3) experts from different disciplines (e.g., economists, political scientists); (4) those with access to classified information and those without; (5) those with prestigious institutional affiliations and those without; (6) those who had lived for lengthy periods in the relevant country and those who had not; (7) those with and without relevant language skills; (8) those who identified their ideology as liberal versus those who considered themselves to be conservative; (9) those who classified themselves as realists (who believe that in world politics, the strong do what they will and the weak accept what they must) versus those who classified themselves as institutionalists (who believe that international institutions have some normative force not reducible to power politics); and (10) those whose temperamental self-identification was boomster-optimist versus doomster-Malthusian. One of my conclusions was

that, in a complex, probabilistic world, we reach the point of diminishing marginal-predictive returns for knowledge considerably more quickly than most experts—and most users of expertise—appreciate.

The findings of Green and Armstrong (2004) also agree with other findings I reported—findings that do pass conventional levels of statistical significance. Green and Armstrong (2004) found that an experimental manipulation that encouraged forecasters to use historical analogies in more sophisticated ways (e.g., a balanced appreciation for key differences and similarities across the range of possible analogies) did produce significant increases in forecasting accuracy. I did not, however, rely on any experimental manipulations of cognitive style; rather, I focused on naturally occurring individual variation among experts in their styles of reasoning. I measured variation both by a cognitive style scale—the hedgehog-Fox scale—and by content analysis of thought protocols that experts generated in support of their predictions. These considerable differences in methodology notwithstanding, I too found evidence that experts who use historical analogies in more flexible and balanced ways (rather than just focusing on the salient points of similarity between the current situation and their favorite analogy) provided significantly more accurate forecasts. Experts who used history predominantly to confirm their hypotheses made predictions that were too extreme. For instance, in 1992, it would have helped experts to be aware that although there were several similarities between North Korea and Romania, there were also many important differences; these differences were sufficient to lessen the subjective probability that the North Korean leadership would be overthrown similar to how the Romanian leadership had been a few years earlier. In 2003, it might have helped to be aware that although there were several similarities between the leadership of Saddam Hussein in Iraq and the Nazi regime of Adolf Hitler, there were also alternative, less ominous, historical analogies, including Italy under Mussolini, the Soviet Union under Stalin, Romania under Ceausescu, Yugoslavia under Tito, and Egypt under Nasser. Using the alternative analogies would have led one to expect a leadership in Iraq that was considerably more risk averse than does the Nazi analogy. That assessment, in turn, might well influence judgments about the subjective likelihood that Iraq would serve as a sponsor for international terrorist strikes against the United States.

Finally, like Green and Armstrong, I find that even the good news about factors that promote forecasting accuracy tends to have some negative aspects: It is hard to raise expert forecasting accuracy appreciably above that possible from simple statistical models. This is a recurring theme in the psychological literature that has, over the last five decades, pitted clinical versus actuarial approaches to prediction against each other (Arkes 2001).

If experts' predictions are as unimpressive as the results of Green and Armstrong and of my own work suggest, why is this fact not more widely appreciated? In politics, one obvious answer is that people are simply too partisan to notice the prediction failures by the pundits on their side—even though they very much savor the prediction failures of opposition pundits. As research on cognitive consistency, performed over several decades, suggests (Abelson et al. 1968), there is some truth to this conjecture.

In closing this commentary, I suggest a more unsettling possibility. Imagine this symbiotic relationship. Experts have an obvious professional self-interest in sustaining the widespread impression that they possess special knowledge about the future and should be frequently consulted. And, as I (1999, 2005) have reported, experts also have an impressive ability to redefine relatively inaccurate forecasts as relatively accurate by invoking belief-system defenses such as "just off on timing" (be patient, x has not happened yet, but it will), the close-call counterfactual (be reasonable, x did not happen but it almost did and would have but for this exogenous shock that no one could have foreseen), and the "I-made-the-right-mistake" (better to have under- or over-estimated them than the opposite mistake). However, many social psychologists have argued, as I have (Tetlock 2005), that people have a deep-rooted need to believe that they live in a predictable and controllable world, and reliance on expert judgment helps to sustain this comforting illusion. Consumers of expertise do not want to believe that in making important decisions—such as whether to go to war or to redirect economic

or trade policy—they could do just as well by relying on simple, extrapolation algorithms or even coin tosses. Each side needs the other too much to disengage from the relationship merely because it is based on an illusion.

References

Abelson, R. P., E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, P. H. Tannenbaum, eds. 1968. Theories of Cognitive Consistency: A Sourcebook. Rand McNally, Chicago, IL.

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

Green, K. C., J. S. Armstrong. 2004. Structured analogies for forecasting. Working Paper 17/04, Monash University, Victoria, Australia. Retrieved November 10, 2005 http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/2004/wp17-04.pdf.

Tetlock, P. E. 1999. Theory driven reasoning about possible pasts and probable futures: Are we prisoners of our perceptions? Amer. J. Political Sci. 43 335-366.

Tetlock, P. E. 2005. Expert Political Judgment: How Good Is It? How Can We Know? Princeton University Press, Princeton, NJ.