Is the Evidence on Forecasting Conflicts Based on Proper Science?

J. Scott Armstrong
The Wharton School, University of Pennsylvania, Philadelphia, PA
November 13, 2001

Abstract

Green's (2002) study on the accuracy of forecasting methods for conflicts does well against traditional scientific criteria. Moreover, it is useful, as it examines actual problems by comparing forecasting methods as they would be used in practice. Some biases exist in the design of the study and they favor game theory. As a result, the accuracy gain of game theory over unaided judgment may be illusory, and the advantage of role playing over game theory is likely to be greater than the 44% error reduction found by Green. The improved accuracy of role playing over game theory was consistent across situations.

In Armstrong (1997a), I reviewed *Co-opetition by* Nalebuff and Brandenburger (1996). Their use of game theory to analyze real-world situations seemed compelling. I concluded that it was unfortunate that the decision makers had not engaged the help of game theorists before they made their decisions. I had some misgivings about the book, however. For example, was there any evidence that game theory had led to better decisions or predictions in conflicts? So I contacted the authors. Brandenburger responded that he was not aware of any studies of the predictive validity of game theory and I was unable to find any such studies.

Many hundreds of academics have been working on game theory for half a century. Thus, it seems strange that it is difficult to find evidence on its predictive validity. Imagine that hundreds of medical researchers spent half a century developing drugs without testing whether they worked as predicted. They would not be allowed to put their drugs on the market.

Green sent me an early draft of his paper in July 2000**.** I thought that it was an important contribution. Here was (1) an important problem that (2) challenged existing beliefs, (3)

contained surprising results, (4) used simple methods, (5) provided full disclosure, and (6) explained it all clearly. In short, Green violated all rules in the "Author's Formula" (Armstrong 1982). That formula, based on a review of empirical research, was updated in Armstrong (1997b). Given the violations, I concluded that reviewers would reject the paper. As a result, with the permission of the Editor of the *International Journal of Forecasting,* Jan deGooijer, I informed the author that his paper would be accepted, subject to reasonable responses to any substantive reviewer concerns that might arise.

Before launching into my analysis, it is worth noting that Green was systematic about his own evaluation of his study. He rated the study on the 32 principles for the evaluation of forecasting methods from Armstrong (2001c). His study did well on 28 of the principles, poor on three, with one being judged as not relevant. I have reviewed these ratings and am in agreement. The rating are summarized at http://decision.co.nz/ratings.pdf.

I discuss whether 1) the problem is important, 2) the findings are important, and 3) the study was done in a scientific manner. I then provide suggestions for further research.

Important Problem?

Green's problem can be stated in two parts: Is it useful to accurately forecast decisions by parties involved in conflicts? If so, which method can best improve upon the way that people currently forecast such decisions?

With respect to the first question, it seems that by better predicting the decisions by the other party in a conflict, one can make better decisions. For example, in 1975, Britain refused to sell the Falkland Islands to a group of Argentine investors backed by the Argentine government. As a result, they had to fight a war, which was clearly a less profitable alternative for Britain as it destroyed resources and people. The three Argentine generals involved did not anticipate Britain's response once Argentine troops occupied the Falkland Islands. Thus, they lost a war and their jobs.

Predictions of decisions might also be of interest to parties outside the conflict. For example, in the case involving the negotiations between the National Football League owners and the Players Association, an insurance company offered strike insurance to the players. To do so, they had to forecast the likelihood that the players would decide to strike.

With respect to the second question, Green examined some of the more important ways that have been recommended for such situations. For example, game theory is often suggested as a way to predict the behavior of rational decision makers, and we have ample evidence from economics that predictions of rational responses are often accurate, even when surprising. Consider for example the predictions of negative consequences by economists for programs to provide economic assistance to single mothers who have young children; many people, including the leading policy makers in some countries, were surprised that people responded to economic incentives in that situation.

I conclude that the problem is important.

Important Findings?

Green's results show substantial differences in accuracy. On average, the best method, role playing, had half the error rate of the worst method, unaided judgment, in predicting actual decisions. Improvements were found in each of the six situations. These findings were obtained from over 1,100 participants. Seldom in studies of forecasting does one encounter such large improvements in accuracy. For example, combining, which is regarded as one of the more important techniques in forecasting, reduces error by about 12% (Armstrong 2001b). I conclude that the findings are important.

Competent Science?

I examined aspects of the scientific method in Green's study. Partly these are standard issues in scientific methods, and partly they are in response to issues raised by reviewers.

Design objective?

Green used the method of multiple hypotheses. I believe this to be one of the most important procedures in striving for objectivity.

Green tried to avoid biases. Due to practical considerations this was not always possible, as discussed below. As it turned out, the design favored game theory relative to unaided judgment and role playing.

Literature review complete and objective?

Green used literature reviews published by others. He also used references in the papers published in key papers. These procedures offer protection against the claim that he might have been biased in his search. However, he also used the SSCI and Internet searches, so there is the potential for bias in screening the papers. Finally, he sent e-mail messages to 474 game theorists to determine whether relevant research might have been overlooked. Given that researchers often become advocates of their approach, this later procedure was biased in favor of finding results favorable to game theory versus the other methods.

Samples of participants large enough?

Some of the reviewers claimed that the study was flawed because the sample of participants was too small. This criticism is unfounded because the participants based their judgments about the behavior of others, not about their own decision in such a case. Such “expert opinion surveys” need only five to 20 experts, depending on such things as the need for precision, level of expertise, and variability of knowledge among the experts (Ashton 1986;

Hogarth 1978; and Libby and Blashfield 1978). Green obtained forecasts from 21 game theory experts.

Samples of participants representative?

Because the experts are assessing the behavior of others, there is no need to have representative experts. Indeed, one would prefer to have the most capable, experienced, and interested experts. Self selection can help in selecting such experts. (For example, studies of survey research have shown that those who are more interested in a topic are more likely to respond; Armstrong & Overton, 1977). All of the game theory experts were self-selected, whereas, role-players and unaided judges were often captive participants in classes. Self-selection would seem to favor game theory.

Most of the unaided judges had little expertise. Since they must draw in part upon their knowledge of similar situations, they would seem to be at a disadvantage relative to the game theorists.

Sample of situations large enough?

Green's study was based on six situations. Additional situations would improve the confidence that one might have in the results. Still, using the Wilcoxon signed-ranks test (one-tail), the probability of getting such results would be only .03 if there were no real difference between the accuracies of game theory and role playing.

In Armstrong (2001a), I suggested that role playing was most appropriate when the interactions in a conflict are examined. Because the Panabla did not involve interactions, I recalculated the percentage of correct responses with it excluded. A striking picture emerges. Chance, unaided judgment, and game theory produce virtually identical results with about 28% correct predictions, compared with 61% for role playing.

To assess the sensitivity of these results to the selection of the situations, I then excluded each of the other situations. This allowed for a comparison of the average accuracy with each combination of the remaining four situations. Again, the results were consistent, as shown in Table 1. In addition, as noted in the last column, the error reductions of role playing over game theory is similar across these analyses. The error reduction is calculated as 100 X (GT"s wrong predictions minus RP's wrong decisions)/ GT's wrong predictions.

Table 1: Average Percentage of Correct Forecasts with Some Situations Excluded

Excluded situations	Chance	Unaided judgment	Game theory	Role playing	Error reduction RP vs. GT
Panalba	28	27	28	61	46
Also excluding:					
Artists	31	32	33	69	54
Distribution	27	32	27	58	42
55%	29	27	27	61	47
Zenith	27	26	29	62	46
Nurses	27	17	22	56	43

Sample of situations representative?

One possibility is that the selection of situations are more likely to include "interesting" cases, and that they might have been considered to be interesting because they were hard to assess judgmentally. This would constitute a bias against unaided judgment, thus favoring game theory and role playing. To assess this, I examined the extent of the error reductions of role playing relative to unaided judgment versus the degree of difficulty for unaided judgment. The results, shown in Table 2, show no evidence of bias with respect to easy versus difficult situations. For example, the error reduction was 48% for the three easiest and 50% for the three most difficult. This analysis does not rule out other sources of bias, however.

Table 2: Were the situations biased against judgment?

Situations	Correct by Judgment	Correct by Role playing	% Error Reduction
Artists' reprieve	5	29	25
Distribution plan	5	75	74
55% plan	27	60	45
Zenith	29	59	42
Panalba	34	76	64
Nurses	68	82	44

Instructions followed?

Green was studying a practical issue. Assume that you have a conflict situation. Does it help to appeal to leading game theorists to make predictions by using game theory? In a real situation, the extent to which game theorists are successful would depend not only on the value of game theory, but on whether analysts can successfully match the situation to their knowledge of game theory, and the extent to which game theory gives them a better understanding of the situation. One would expect that those with more experience in game theory would be more skillful at applying game theory to these situations. However, Green found that those with more experience in game theory were no more accurate in their predictions. He also found that those spending more time were not more likely to be correct.

The instructions for unaided judgment were easy to follow. However, the role playing was done by people who had little prior experience with this approach. Most were students, so it seems likely that some might not have taken the exercise seriously. As a result, game theory had an advantage over role playing.

External incentives provided to the participants were minimal. Would the result have changed had there been financial incentives? Remus, O'Connor and Griggs (1998) examined the evidence on this issue. Based on ten studies, they conclude that there is little evidence that financial incentives would improve accuracy for judgmental forecasting

Intrinsic incentives would seem to favor the game theorists, as they were asked to use game theory in making predictions. Presumably, they would want to see game theory do well, whereas the other participants had no attachment to their method.

Biased administration?

The experiments were conducted by those who had a prior hypothesis. Might this have produced an unintended bias that might lead the participants to act as the researchers expected? This has been referred to as "demand effects." Sigall, Aronson and van Hoose (1970), examined the evidence and found little support for the theory that participants are cooperative with the experimenters. Rather, their concern seems to be to present themselves in a favorable light, which would suggest a bias in favor of game theorists.

As with any study, bias might occur for other reasons. Thus, it would be useful if the studies could be replicated or extended by those who might believe that game theory is superior to role playing.

Biases from variations in administration?

One reviewer claimed that the design was faulty because there were variations in the administrative procedures. For example, different times were allowed for different administrations. In my opinion, variations are useful when there is a potential for bias. Thus, for example, researchers are typically advised to vary the order of the presentation of materials to participants (as Green did). Variations also allow one to assess whether administrative procedures have any effect. On the whole, I saw the variations as a benefit to Green's design.

Full disclosure?

Green reported on all of his procedures. Some of the details are provided on the Internet. This includes information about the participants and explanations of how they made their

predictions. For description of the reasoning used by the game theorists, see decision.co.nz/approach.pdf

He was responsive to reviewers when they asked for additional explanation. As nearly as I can judge, he has met the requirements for full disclosure.

Other criteria examined?

The use of a forecasting method also depends on the cost, acceptability, and other factors. Green provides some details on costs and game theory was the most expensive approach. As for acceptability, it seems reasonable to hypothesize that role playing, by showing a vivid and detailed prediction of the decisions, would be compelling to decision makers. It would be useful to make empirical comparisons on the acceptability of unaided judgment, game theory, and role playing forecasts.

Green's study focused primarily on predictive validity. Game theory may have other uses such as improving the search for alternative solutions although I expect that formal idea generation procedures, such as brainstorming, will prove superior to game theory. Can game theory substantially improve the way managers think about problems in comparison, say, to someone instructed to calculate net present values for alternatives? In short, I have been unable to find evidence that game theory has any practical value for managers.

Clearly written?

An important aspect of good research is that it be clearly written. Green's paper has a Flesch-Kinkaid readability index equal to 12th grade. This is much more readable than typical scientific papers.

Further Research

While a single study is vastly superior to no studies, it cannot be expected to resolve all of the issues. To date, the level of research effort devoted to game theory is thousands of times that devoted to alternative procedures for analyzing conflicts. My primary recommendation is that game theorists should adopt the method of multiple hypotheses such that it embraces procedures other than game theory.

Does game theory add to an analyst's way of making predictions in real situations? As noted above, the procedures were biased in favor of game theory. What if 474 non-game theorist adults of similar background as the game theorists were contacted, and the same situations were presented. Assume then that those most interested made unaided judgmental predictions. Would their accuracy be equal to that provided by the game theorists? If so, one could conclude that the superiority of the game theorists in Green's study was due to their experience rather than to their knowledge of game theory.

Conflicts vary, so it would be useful to study the conditions under which each approach is most effective. This would be aided by examining more situations, especially if substantially different from those in Green's study. Note, for example, that game theory was better than role playing for the one situation that did not have active interactions between two groups. Goodwin (2002) discusses various types of situations. It would be useful to study real situations that have been suggested by game theorists. To date, however, my appeals to game theorists to supply such situations have gone unanswered.

Experts who have experience with conflict situations might be able to make good forecasts at a lower cost than role playing. This seems especially likely if they identify analogous situations in a structured manner. Research on analogies might help to determine whether it is possible to identify relevant experts, how one should structure the forecasting task, and whether analogies can lead to low-cost predictions that are as accurate as those by role playing.

Green's study has focused on forecasting. One might extend the study to decision making. For example, has game theory led to better decisions than those that could be obtained by other methods, such as evaluating the net present value of alternative strategies? Can game theory produce a better set of strategies than one might achieve by using brainstorming or other creative techniques? To date, despite the enormous efforts that researchers have devoted to research on game theory, I have been unable to find evidence that game theory will improve decision making.

Conclusions

An examination of threats to validity supports Green's comparative study of forecasting methods. The game theorists' predictions were slightly more accurate than those from the use of unaided judgment, though the advantage might have been due to biases in the design, such as the lack of experience on the part of the participants using unaided judgment. Role playing was substantially more accurate than game theory despite biases favoring game theory.

References

Armstrong, J. S. (2001a), "Role playing: A method to forecast decisions," in J. S. Armstrong (ed.): *Principles of Forecasting: A Handbook for Researchers and Practitioners.* Norwell, MA: Kluwer Academic Publishers

Armstrong, J. S. (2001b), "Combining forecasts," in J. S. Armstrong (ed.): *Principles of Forecasting: A Handbook for Researchers and Practitioners.* Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (2001c), "Evaluating methods," in J. S. Armstrong (ed.): *Principles of Forecasting: A Handbook for Researchers and Practitioners.* Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (1982), "Barriers to scientific contributions: The author's formula," *Behavioral and Brain Sciences,* 5 (June), 197-199.

Armstrong, J. S. (1997a), "Why can't a game be more like a business? A review of *Co-opetition* by Brandenburger and Nalebuff," *Journal of Marketing*, 61 (April), 92-95.

Armstrong, J. S. (1997b), "Peer review for journals: Evidence on quality control, fairness, and innovation," *Science and Engineering Ethics,* 3, 63-84.

Armstrong, J. S. & T. S. Overton (1977), "Estimating nonresponse bias in mail surveys," *Journal of Marketing Research*, 14, 396-402.

Ashton, A. H. (1986), "Combining the judgments of experts: How many and which ones?" *Organizational Behavior and Human Decision Processes,* 38, 405-414.

Brandenburger, A. M. & B. J. Nalebuff (1996), *Co-opetition.* New York: Doubleday.

Goodwin, P. (2002), "Forecasting games: Can game theory win?" *International Journal of Forecasting* (forthcoming)

Green, K. C. (2002), "Forecasting decisions in conflict situations: A comparison of game theory, role playing and unaided judgment," *International Journal of Forecasting* (forthcoming)

Hogarth, R. M. (1978), "A note on aggregating opinions," *Organizational Behavior and Human Performance*, 21, 40-46.

Libby, R. & R. K. Blashfield (1978), "Performance of a composite as a function of the number of judges," *Organizational Behavior and Human Performance*, 21, 121-129.

Remus, W., M. O'Connor & K. Griggs (1998), "The impact of incentives on the accuracy of subjects in judgmental forecasting experiments," *International Journal of Forecasting*, 14, 515-522.

Rowe, G. & G. Wright (2001), "Expert opinions in forecasting: Role of the Delphi technique," in J. S. Armstrong (ed.), *Principles of Forecasting.* Norwell, MA: Kluwer Academic Press.

Sigall, H., E. Aronson and T. van Hoose (1970), "The cooperative subject: Myth or reality," *Journal of Experimental Social Psychology*, 6, 1-10.