International Journal of Forecasting 18 (2002) 345–352

www.elsevier.com/locate/ijforecast

Assessing game theory, role playing, and unaided judgment

J. Scott Armstrong

The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

Green's study [*Int. J. Forecasting* (forthcoming)] 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. For those cases that simulated interactions among people with conflicting roles, game theory was no better than chance (28% correct), whereas role-playing was correct in 61% of the predictions.

Keywords: Forecasting; Role playing; Simulated interactions

1. Introduction

In Armstrong (1997a), I reviewed *Co-opetition* by Brandenburger and Nalebuff (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 finding evidence on its predictive validity is difficult. 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 market their drugs.

Kesten Green sent me an early draft of his paper in July 2000 (Green, 2002). I thought it was an important contribution because he: (1) described an important problem, (2) challenged existing beliefs, (3) obtained surprising results, (4) used simple methods, (5) provided full disclosure, and (6) explained it all clearly. In

E-mail address: armstrong@wharton.upenn.edu (J.S. Armstrong).

0169-2070/02/$ – see front matter
PII: S0169-2070(02)00024-9

short, Green violated all the rules in the 'Author's Formula' (Armstrong, 1982). That formula, based on a review of empirical research, was updated in Armstrong (1997b). Given Green's violations of the formula, I expected that reviewers would reject the paper. To avoid rejection, with the permission of Jan deGooijer, Editor of the *International Journal of Forecasting*, I informed Green that his paper would be accepted, subject to reasonable responses to any substantive reviewers' concerns.

Green had been systematic in 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, poorly on three, with one judged as not relevant. I have reviewed these ratings and am in agreement. The ratings are at kestencgreen.com/ratings.pdf.

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

2. Important problem?

Green's problem can be stated in two parts: Is it useful to accurately forecast the decisions made by parties in conflict? 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 of one's adversary 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, it had to fight a war to defend its ownership, which was clearly a less profitable alternative for Britain than selling the islands. The three Argentine generals involved had not anticipated Britain's response to Argentinian troops occupying the Falkland Islands. They lost the war and their jobs.

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

With respect to the best method to use, Green examined some of the more important methods 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.

The problem of predicting decisions in conflict situations is important.

3. Important findings?

Green's results show substantial differences in accuracy among methods. On average, the best method, role playing, had half the error rate of the worst method, unaided judgment, in predicting actual decisions. In five of the six situations, he found that role playing improved accuracy over other methods. These findings were obtained using over 1100 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). Green's findings are important.

4. Competent science?

I examined Green's use of the scientific method. Considering standard issues regarding scientific methods and issues raised by reviewers.

4.1. Was the design objective?

Green used the method of multiple hypotheses. I believe this is an important procedure in striving for objectivity.

Green tried to avoid biases. Because of practical considerations, he could not always do so. As it turned out, the design of his study favored game theory relative to unaided judgment and role playing.

4.2. Was the literature review complete and objective?

Green used literature reviews published by others and references listed in key papers. These procedures offer protection against the claim that he might have been biased in his search. However, he also used the *Social Science Citation Index* and Internet searches, which might lead to bias when 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 advocate their own approaches, consulting game theorists might have produced findings favorable to game theory.

4.3. Were the 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 predictions, not on their own behavior, but on their knowledge about the behavior of many people. As a result, 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; Libby & Blashfield, 1978). Green obtained forecasts from 21 experts in game theory.

4.4. Were the samples of participants representative?

Because Green's experts assessed the behavior of others, there was no need to have representative experts. Indeed, one would prefer the most capable, experienced, and interested experts. Thus, self-selection is beneficial. (Studies of survey research have shown that people who are more interested in a topic are more likely to respond; Armstrong & Overton, 1977). All of the game theory experts Green used were self-selected, whereas the role-players and unaided judges were often captive participants in classes. Self-selection would favor game theory here.

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

4.5. Was the sample of situations large enough?

Green's study was based on six situations. Additional situations would improve one's confidence in the results. Still, using the Wilcoxon signed-ranks test (one-tail), the probability of getting such results would be only 0.03 if game theory and role playing were equally accurate methods.

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 excluded it and recalculated the percentage of correct responses. A striking picture emerged. 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 situations, I then excluded each of the other five situations, one at a time. This allowed for a comparison of the average accuracy with each combination of the remaining four situations. Again the results were consistent (Table 1). The error reductions of role playing over game theory were similar across these analyses.

Given these results, rather than using the awkward term ‘‘role playing that simulates interactions among people’’, we might use the term ‘‘simulated interactions’’.

4.6. Was the sample of situations representative?

One possibility is that when selecting situations one is more likely to include ‘interesting’ cases, and that one might have considered them 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 possibility, I examined the extent of the error reductions of role playing relative to unaided judgment and compared this to the extent to which the correct answers were obvious to unaided judges. The results (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.

Table 2
Were the situations biased against judgment?

Situations	Correct by judgment	Correct by role playing	Percent error Reduction[a]
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

[a] The error reduction was calculated as 100 × (unaided judgments’ wrong predictions minus RP’s wrong decisions)/(unaided judgments’ wrong predictions).

4.7. Did the participants follow the instructions?

Green studied a practical issue. Assume that you have a conflict situation. Would it help you to ask leading game theorists to make predictions and to also ask that they make use of any relevant expertise in game theory? In a real situation, the extent to which game theorists

Table 1
Average percentage of correct forecasts with some situations excluded

Excluded situations	Chance	Unaided judgment	Game theory (GT)	Role playing (RP)	% Error reduction[a] 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

[a] Error reduction was calculated as 100 × (GT’s percent wrong predictions minus RP’s percent wrong decisions)/(GT’s percent wrong predictions).

were successful would depend not only on the value of game theory, but on whether they could successfully match the situation to their knowledge of game theory, and the extent to which game theory gave 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 than novices. He also found that those spending more time were not more accurate.

The instructions for unaided judgment were easy to follow. However, the subjects using role playing had little 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 in that the game theorists should have been able to follow the instructions.

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

Intrinsic incentives would seem to favor the game theorists because 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 methods.

4.8. Was the administration biased?

The experiments were conducted by researchers who had a prior hypothesis. Might this have produced an unintended bias that led participants to act as the researchers expected? Such effects are called 'demand effects.' Sigall, Aronson, and van Hoose (1970) examined the evidence and found little support for the theory that participants cooperate with experimenters. Rather, their concern seems to be to present themselves in a favorable light. Because the game theorists all identified themselves, this should have led them to have more of a concern about looking good.

As with any study, bias might occur for other reasons. Thus, it would be useful if someone who believed that game theory might have some advantages over role playing would replicate or extend the studies.

4.9. Were there biases from variations in administration?

One reviewer claimed that the design was faulty because there were variations in the administrative procedures. For example, Green allowed different times for different administrations of the forecasting methods. In my opinion, variations are useful when a potential for bias exists. 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.

4.10. Did Green provide full disclosure?

Green reported on all of his procedures, providing important details on the Internet. He included information about the participants and explanations of how they made their predictions. For a description of the reasoning the game theorists used, see kestencgreen.com/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.

4.11. Would the use of other criteria affect the conclusions?

Green's study focused primarily on predictive validity. The use of a forecasting method also depends on its cost, acceptability, assessment of risk, and other factors. Green provides some details on costs; game theory was the most expensive approach.

It would be useful to make empirical comparisons on the acceptability of unaided judgment, game theory, and role-playing forecasts. It seems reasonable to hypothesize that role playing, by showing a vivid and detailed prediction of decisions, would be compelling to decision makers.

It is not clear how game theory would alert decision makers to risk. In contrast, risk can be assessed through unaided judgment and role playing. Consider, for example our study of the journal royalties negotiation, which involved the International Institute of Forecasters and John Wiley Publishers (Armstrong & Hutcherson, 1989). Unaided judgment suggested that there was a 12% chance that the negotiations would lead to a cancellation of the contract, and role playing predicted a 42% chance of cancellation. Cancellation would produce substantial losses to both parties. Thus, even though the prediction was that there would be an agreement on royalties, it would have been prudent to take steps to avoid a cancellation. In fact, the actual outcome was a cancellation of the contract.

Game theory may have uses other than forecasting, such as improving the search for alternative solutions, although I expect that formal idea-generation procedures, such as brainstorming, would prove superior to game theory for that purpose. Can game theory substantially improve the way managers think about problems compared to, say, calculating net present values for alternatives?

In general, then, Green's results on the superiority of role playing hold up for a variety of criteria.

4.12. Was the paper 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. It is much more readable than typical scientific papers.

5. Further research

Green's is the first study on the predictive validity of game theory. While a single study is superior to having no studies, it cannot resolve all of the issues. To date, the research effort devoted to game theory is probably 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 and embrace 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 with backgrounds similar to those of the game theorists were contacted, and the same situations were presented? Assume then that those most interested made unaided judgmental predictions. Would they be as accurate as 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. To do this, one could examine more situations, especially if they differed substantially from those in Green's study. Note, for example, that game theory was better than role playing for the one situation that did not simulate the interactions between groups. Goodwin (2002) discusses various types of situations. It would be useful to study real situations that were 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. They would be especially likely to do so if they identify analogous situations in a structured manner. Research on analogies might help one to determine whether it is possible to identify relevant experts, how one should structure the forecasting task, and whether analogies could lead to low-cost predictions that were as accurate as those by made by role playing.

In his study, Green 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 using game theory produce a better set of strategies than using brainstorming or other creative techniques? To date, despite the enormous efforts devoted to research on game theory, I have been unable to find evidence that game theory will improve decision making.

6. Conclusions

The game theorists' predictions were slightly more accurate than those from participants using unaided judgment, although the advantage may have arisen from biases in the design, such as the lack of experience on the part of the participants using unaided judgment. This advantage does not apply however, when the situations are restricted to conflicts involving a series of interactions among the parties in conflict. Role playing was substantially more accurate than game theory despite biases favoring game theory.

In general, an examination of the procedures used in Green's comparative study of forecasting methods supports his findings. That said, much can be learned from further study of these issues. Hopefully, game theory researchers (and others) will conduct empirical studies that would assess the value of game theory relative to other approaches. More important, despite the substantial benefits identified in research to date, little research has been done on the use of role playing as a forecasting technique. In particular, research is needed for simulated interactions in cases involving conflicts. We know little about how to best implement role playing and about the conditions under which it is most effective relative to other methods.

References

Armstrong, J. S. (1982). Barriers to scientific contributions: the author's formula. *Behavioral and Brain Sciences*, *5*, 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*, 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. (2001a). Role playing: a method to forecast decisions. In Armstrong, J. S. (Ed.), *Principles of forecasting: a handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (2001b). Combining forecasts. In Armstrong, J. S. (Ed.), *Principles of forecasting*. Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (2001c). Evaluating methods. In Armstrong, J. S. (Ed.), *Principles of forecasting*. Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S., & Hutcherson, P. D. (1989). Predicting the outcome of marketing negotiations. *International Journal of Research in Marketing*, *6*, 227–239.

Armstrong, J. S., & Overton, T. S. (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., & Nalebuff, B. J. (1996). *Co-opetition*. New York: Doubleday.

Goodwin, P. (2002). Forecasting games: can game theory win? *International Journal of Forecasting*, *18*, 369–374.

Green, K. C. (2002). Forecasting decisions in conflict situations: a comparison of game theory, role playing and unaided judgment. *International Journal of Forecasting*, *18*, 321–344.

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

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

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

Sigall, H., Aronson, E., & van Hoose, T. (1970). The cooperative subject: myth or reality. *Journal of Experimental Social Psychology*, *6*, 1–10.