DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125

ON ELICITING BELIEFS IN STRATEGIC GAMES

Thomas R. Palfrey
California Institute of Technology

Stephanie W. Wang
Princeton University

SOCIAL SCIENCE WORKING PAPER 1271

July 2007

On Eliciting Beliefs in Strategic Games[1]

Thomas R. Palfrey[2] and Stephanie W. Wang[3]

March 2007

[1]We gratefully acknowledge the financial support of the National Science Foundation (SES-0079301, SES-0450712, SES-0094800), The Princeton Laboratory for Experimental Social Science, and the Center for Economic Policy Studies. We are grateful for comments from Juan Carrillo.

[2]The Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125.

[3]Department of Economics, Princeton University, Princeton, NJ 08540.

Abstract

Recent experimental studies have used scoring rules to measure beliefs of subjects engaged in strategic games with other subjects. Using data from one such study, we conduct a series of experiments where our experienced subjects observe early rounds of data from that study and are given monetary incentives to report forecasts of choices in later rounds. We elicit beliefs using three different scoring rules: linear, logarithmic, and quadratic. There are significant differences between the elicited beliefs under quadratic and logarithmic scoring rules in spite of both being proper scoring rules. The linear scoring rule does not always elicit degenerate priors as theory predicts, but does result in reported beliefs that are closer to 0 and 1 than the proper rules. We also compare the forecasts of our trained observers to elicited beliefs of the actual players in the Nyarko-Schotter experiment and find significant differences. There was a significant positive correlation between observer forecasts and the choice behavior in the game under all three of our scoring rules while there was no significant correlation between the players' own forecasts and the actual play. This raises doubts about whether beliefs can be reliably elicited from players who simultaneously have a stake in the target of their forecast, in this case the opponent's choice. The distribution of forecasts by NS players also had more dispersion than the observer forecasts using either of the proper scoring rules, but slightly less dispersion than the observer forecasts using the linear scoring rule. We also find evidence of both belief convergence and information aggregation when beliefs are elicited iteratively from a group.

Keywords: Scoring rules; Experiments; Game theory; Forecasting; Beliefs; Information aggregation

1 Introduction

Probabilistic beliefs play a central role in mathematical theories of strategic decision making. In games of strategy, optimal decisions depend on beliefs about other players' choices, which in turn depend on their beliefs about one's own decision, and so on. Many ideas lying at the very foundation of these theories and related concepts in economics, such as rational expectations and Nash equilibrium are built around strong assumptions about beliefs. Most attempts to test these theories, often in laboratory experiments, either measure beliefs indirectly by estimation, or impose maintained hypotheses about beliefs (such as rational expectations), resulting in tests of joint hypotheses about beliefs and rational choice. The ability to evaluate or test these theories more sharply would be greatly enhanced if it were possible to measure beliefs directly. Indeed, a number of recent attempts of direct measurement of probabilistic beliefs have been attempted by experimental economists, in the context of strategic games. Examples include Dominitz and Hung (2004), in the context of information cascades, Huck and Weizsacker (2002) and Offerman et al. (2006) in the context of lottery choice experiments, McKelvey and Page (1990) for information aggregation, Duwfenberg and Gneezy (2000) in trust games, Offerman et al. (1996) and Croson (2000) in voluntary contribution games,The results of those papers raise questions about the measurement methodology itself, and its applicability to the elicitation of beliefs in a strategic environment. Indeed, a striking finding from several of these experiments is the surprising prevalence of extreme forecasts (degenerate or nearly degenerate forecasts), which is hard to reconcile with standard theory.

This paper explores four methodological questions and two substantive questions about the use of scoring rules for the elicitation of probabilistic beliefs about behavior in strategic games. We undertake this exploration in the context of a simple 2x2 asymmetric matching pennies game similar to the one originally studied by Ochs (1995)

and more recently by McKelvey, Palfrey, and Weber (2000), Goeree, Holt, and Palfrey (2003), and Nyarko and Schotter (NS, 2002).

The first question is: can beliefs be reliably elicited from the players of a game, during the play of the game? Unreliable reported beliefs could arise for a variety of reasons, including psychological factors such as rationalization, or via distortion of incentives because they are also being paid according their play in the game, which violates the "no-stakes" condition of Kadane and Winkler (1988). We address this question by comparing the elicited beliefs of (experienced) observers to the elicited beliefs of the players themselves. Our subjects observe real sequences of choice behavior from the NS data, and are asked to make probabilistic one-move-ahead forecasts of the play of the game, as the sequence is played back to them in real time, using scoring rules to incentivize the forecasts. Because the NS subjects also made incentivized one-move-ahead forecasts, this allows for a direct comparison.

The other three methodological questions address the issue of whether the choice of the scoring rule makes a difference: Are forecasts elicited using proper scoring rules systematically different from those elicited using improper scoring rules? Are forecasts elicited via two different proper scoring rules the same or different? Are forecasts better calibrated for some scoring rules than others? With these latter two questions in mind, we conduct an experiment with three different treatments, each corresponding to a different scoring rule. The three scoring rules used are *logarithmic* (proper), *quadratic* (proper), and *linear* (improper).

The substantive questions both concern information aggregation and belief convergence of subjective beliefs. First, are individuals in a group able update their beliefs in response to the forecasts of other members of the group? (*belief convergence*) Second, if such convergence occurs, are individual forecasts improved by group interaction? (*information aggregation*) To address these questions, our experiment

includes a second feature that allows for information aggregation. Our observers were placed in groups of four, and there were two sequential rounds for each forecast. The entire profile of individual forecasts of group members was revealed between the two rounds, so each individual had an opportunity to update his or her forecast in response to the forecasts of the other group members. This allows us to test for belief convergence (comparing the variance of first round to second round forecasts) and information aggregation (comparing the accuracy of first round and second round forecasts).

We have five main findings. First, there is a significant difference between the elicited beliefs under quadratic and logarithmic scoring rules in spite of both being proper scoring rules. Forecasts elicited by the logarithmic scoring rule have less dispersion (closer to (.5,.5)), and are better calibrated than forecasts under the quadratic scoring rule according to several different measures. Second, the linear scoring rule does not always elicit degenerate priors, but does result in reported beliefs that are closer to 0 and 1 than the proper rules, and these forecasts are more poorly calibrated. Third, the forecasts by our observers were more accurate than the forecasts of the NS players, in the sense that the average elicited forecast was closer to the true choice frequencies in the data for all three scoring rules. Furthermore, there was a significant positive correlation between observer forecasts and the choice behavior in the game for all three of our scoring rules while there was no significant correlation between the players' forecasts and the actual play. This reinforces doubts about whether beliefs can be reliably elicited from players who simultaneously have a stake in the target of their forecast, in this case his opponent's choice. Fourth, the distribution of forecasts by NS players had more dispersion than the observer forecasts using either of the proper scoring rules, but slightly less dispersion than the observer forecasts using the linear scoring rule. Fifth, we find evidence of both belief convergence and information aggregation.

1.1 Related Literature

1.1.1 Scoring rules

Scoring rules, which yield payoffs as a function of vector of probabilistic forecasts and a realized event, are used to elicit subjective probabilities in laboratory and real-life settings. Different scoring rules have different incentive compatibility properties. Because elicitation methods are used to uncover "true" probabilistic beliefs, incentive compatibility is an important criterion for the "goodness" of any scoring rule. A scoring rule is classified as proper if it is incentive compatible. In the scoring rule literature, a scoring rule is considered incentive compatible if a forecaster cannot attain a higher expected score by reporting a probability different than her true probability.

Brier (1950) and Good (1952) were the first to identify two such proper scoring rules, *quadratic* and *logarithmic,* respectively. Since then, both the quadratic and logarithmic scoring rules as well as another well-known one, the spherical scoring rule, have been shown to be strictly proper (Winkler and Murphy 1968), meaning that the expected utility is uniquely maximized when the stated probabilities are equal to the true probabilities. Savage (1971) specifies the general rule for generating the class of strictly proper scoring rules and there have been numerous comparative studies of desirable and undesirable properties of proper scoring rules such as quadratic, logarithmic, spherical, ranked probability, and utility. (De Finetti 1965; Roberts 1965; Murphy 1969; Winkler 1969; Staël von Holstein 1970). Kadane and Winkler (1988) have studied the effect of risk aversion on forecasts under the quadratic scoring rule. Offerman et. al. (2006) and Fountain (2002) propose procedures to adjust for non-neutral risk attitudes.

1.1.2 Previous experimental results about scoring rules

The quadratic scoring rule is the most common one used in both laboratory and field experimental settings for the forecasting of subjective events such as weather forecasting

(Staël von Holstein 1971), stock market prices (Staël von Holstein 1972), and outcomes of sporting competitions (Winkler 1971), and game theory (see below). The logarithmic scoring rule has been used to a much lesser extent in experiments on education testing (Hambleton et. al. 1970; Glein and Wallace 1974) and information aggregation (Ledyard et. al. 2005).

A few articles have examined whether the ranking of forecasters remain consistent under different scoring rules (Staël von Holstein 1971) by eliciting the subjective probabilities using one scoring rule (quadratic) and calculating the scores under other proper scoring rules (logarithmic, spherical, and ranked probability) with those subjective probabilities. One study (Nelson and Bessler 1989) compares the number of extreme probabilities elicited under a proper scoring rule (quadratic) vs. an improper one (linear). Phillips and Edwards (1966) compare the accuracy and improvement of the subjects' Bayesian inference under several scoring rules.

Existing experimental results about the use of scoring rules to elicit subjective beliefs about action choices in a strategic game are mixed. In the context of two person matrix games, extreme reported beliefs (beliefs with 25% or less likelihood for one of the states) are observed with surprising frequency (Dominitz and Hung 2004, Nyarko and Schotter 2002). Because the "true" frequencies of target states is generally between .35 and .65 this indicates inaccuracies in the forecasts. Furthermore, beliefs are erratic, in the sense that they change much faster from period to period than a Bayesian model would predict, indicating that forecasts are not only inaccurate, but highly imprecise (Nyarko and Schotter 2002, fig. 2, p. 980). If the players were adjusting beliefs according to Bayes rule or even according to a simple counting procedure, truthful reporting of beliefs would have a smoother trajectory than what was actually observed. There is also evidence from two person laboratory games that the process by which subjects decide on a forecast is qualitatively different from the decision process they use

to make a decision, which can sometimes result in forecasts that are inconsistent with choice behavior (Costa-Gomes and Weizsacker 2006).

In contrast, Dominitz and Hung (2004), in the context of an information cascade experiment, report that players' forecasts are dampened relative to Bayesian reports. In particular, they find that subjects often fail to change their forecasts in response to hard information, which suggests distortions in the elicitation procedure. The task was much different from the our task of one-step-ahead forecasts of choices in a repeated game, since their subjects were repeated forecasting a static target (the state of the world), rather than a stochastically moving target. Offerman et al. (1996) elicited subjective probabilistic beliefs about the level of contributions of other players in a voluntary contributions game. Some of the forecasts were degenerate, bimodal, or implausible for other reasons, and they confirm the finding reported by Palfrey and Rosenthal (1991) that subject beliefs about others' contributions are biased upward. Nevertheless, the authors conclude that expectations appear to be reasonable.

There is also mixed evidence about the similarities and differences between forecasts elicited from observers and forecasts elicited from players themselves. Huck and Weizsacker (2002) elicit forecasts from subjects who observe decision makers in a simple (objective) binary lottery choice task. They find some inaccuracies, notably that the forecasts are closer to 50/50 than the actual choice frequencies of the subjects, and that this doesn't depend in a significant way on the elicitation procedure. This is in stark contrast to the forecasting behavior measured using an identical quadratic scoring rule in the NS experiment, where reported beliefs of players are biased in the opposite direction. The combination of these two findings is completely at odds with findings reported in Offerman et al. (1996, p. 828), where observers submitted forecasts that were *more dispersed* than those submitted by the players themselves.

1.1.3 Convergence of beliefs

Our iterative elicitation method could induce a common knowledge inference process whereby individual beliefs adjust after others' beliefs are revealed. In the common knowledge literature, Aumann (1976) first established that if two agents have the same common prior, their posterior probability of an event must be the same if the posteriors are common knowledge. The subsequent work of Geanakoplos and Polemarchakis (1982), McKelvey and Page (1986), and Nielsen et al. (1990) are more closely related to the possible process generated by our iterative elicitation method. Geanakoplos and Polemarchakis show that with iterated exchange of information between the agents, the inference process would terminate at a point where the posterior probabilities are equal. McKelvey and Page demonstrate that for the case with n agents and public knowledge of a summary statistic, the iterated reactions to the public statistic would lead to consensus and complete pooling of private information held by the agents. Winkler (1968) takes a different approach and compares two mathematical methods of combining one-shot individual forecasts into an aggregate one without any informative interaction between the forecasters. He considers the effect of the weighted-average method and the combining natural-conjugate method as well as the choice of weights under both on the aggregate probabilities.

Related to our iterative elicitation method are experiments in which subjects receive feedback about other subjects' forecasts (McKelvey and Page 1990; Offerman and Sonnemans 1998; Choi et. al. 2005; Winkler 1968). With the exception of Winkler's experiment in which he elicits forecasts about subjects with intrinsic uncertainty such as the weather or sports through a questionnaire, the rest induced differences in private information in the laboratory and focused upon the efficiency of private information pooling when there is objective uncertainty. These studies report some belief convergence as measured by the reported forecasts of these objective events.

2 Theoretical Background

2.1 Simple Matrix Game

This is the simple matrix game that was used in the Nyarko-Schotter experiment and in ours as well.

	Green	*Red*
Green	6, 2	3, 5
Red	3, 5	5, 3

Table 1. Matrix game payoffs.

This is a constant sum game with an unique Nash equilibrium in mixed strategies that is supported by the principle of best responses and Nash equilibrium. In equilibrium both players choose Green with 40% probability and Red with 60% probability.

2.2 Three Scoring Rules

Scoring rules, which compute a numerical score as a function of the stated probabilities as well as the realized event, are often used in forecasting and experimental settings to assess the accuracy of forecasts. In our experiment, this score also specifies the monetary payoff. A scoring rule is proper if the subject maximizes her expected monetary payoff by revealing her true belief. We first characterize the three scoring rules used in the three belief elicitation treatments of our experiment. We then go on to show that the quadratic and logarithmic scoring rules are proper whilst the linear scoring rule is not.

2.2.1 Preliminaries

Let $i = 1, 2, ..., n$ denote the n possible events and let $p = (p_1, p_2, ..., p_n)$ be the subject's stated probability distribution where p_i is the stated probability of event i. Define the scoring rule $S = \{S_1, S_2, ..., S_n\}$ as a collection of scoring functions where $S_i(p)$ specifies the score when event i is realized as a function of the subject's stated probability

distribution p. Let $\pi = (\pi_1, \pi_2, ..., \pi_n)$ be the subject's true probability distribution where π_i is the true probability of event i. I_i is an indicator function that takes the value 1 if the realized event is event i and 0 otherwise.

2.2.2 Characterization

1. *Quadratic Scoring Rule:*

$$S_i(p) = \alpha - \beta \sum_{k=1}^{n} (I_k - p_k)^2 \tag{1}$$

where $\alpha, \beta > 0$

The quadratic rule scores the inaccuracy of the forecast as the sum of the square deviations and deducts it from the full score. This deviation is the probability placed on those events that were not realized as well as the probability not placed on the event that was realized. We will now focus on the specific case of the general quadratic scoring rule that pertains to our belief elicitation experiment. There only two possible events, the event that a player chooses Green, which we denote as G, and the event that a player chooses Red, R. In our experiment as well as in the Nyarko-Schotter experiment, α is chosen to be 1 and β is chosen to be 0.5. The score if event G occurs is therefore:

$$\begin{aligned} S_G &= 1 - 0.5((1 - p_G)^2 + (0 - p_R)^2) \\ &= 1 - (1 - p_G)^2 \\ &= 1 - p_R^2 \end{aligned}$$

Similarly, the score if event R occurs:

$$\begin{aligned} S_R &= 1 - 0.5((1 - p_R)^2 + (0 - p_G)^2) \\ &= 1 - p_G^2 \\ &= 1 - (1 - p_R)^2 \end{aligned}$$

The worst that a subject can do is to place the maximum probability on the event that is not realized and the best she can do is to place that same probability on the realized event. Under this quadratic specification, her score can never be negative.

2. *Logarithmic Scoring Rule:*

$$S_i(p) = \alpha + \beta(logp_i) \tag{2}$$

where $\alpha, \beta > 0$

The logarithmic rule deducts for inaccuracy by adding the natural log of the probability placed on the realized event (a negative number since $0 \leq p_i \leq 1$) from the base score. The less the probability placed on the realized event, the greater is the deduction. The score is $-\infty$, however, when zero probability is placed on the realized event. Because of this property, implementation in practice requires bounding the forecasts away from zero. We place bounds on the maximum (0.9) and minimum (0.1) probability a subject can place on any event. We set α to be 1 and β to be 0.45. The score if event G occurs in the logarithmic treatment:

$$S_G(p) = 1 + 0.45(logp_G)$$

The corresponding score if event R occurs:

$$S_R(p) = 1 + 0.45(logp_R)$$

Under this specification, the subject would only receive a negative score (-0.036) if she places the maximum probability on the event that does not occur.

3. *Linear Scoring Rule:*

$$S_i(p) = \beta(p_i) \tag{3}$$

where β>0

The linear scoring rule simply gives the probability placed on the realized event as the score. In the linear treatment, we set β to be 1 so that the prediction score when G is realized is simply p_G and likewise it is p_R when R is realized.

2.2.3 Properties

Selten (1998) specified four axioms and asked which axioms each of the three scoring rules satisfy. The symmetry axiom simply requires that scores should not be dependent upon the labeling of the events. The elongation invariance axiom states that the scores for the realization of each possible event should not be altered by the addition of an impossible event. The incentive compatibility axiom is equivalent to the properness condition we defined above. This is the property of scoring rules that has been scrutinized most intently for obvious reasons. Lastly, the neutrality axiom requires that the score does not favor any particular event in deducting for inaccuracy. For example, in the case of only two events, G and R, placing maximum probability on R when G actually occurs should result in the same expected score loss as placing maximum probability on G when R occurs.

1. *Quadratic Scoring Rule:* Selten shows that the quadratic scoring rule is the only one out of the three to satisfy all four axioms. In particular, it is the only scoring rule that satisfies the neutrality axiom. We will focus on the incentive compatible axiom in the relevant two-event case. The subject's expected payoff given her true belief π:

$$\begin{aligned} E(PO) &= \pi_G(1-p_R^2) + \pi_R(1-p_G^2) \\ &= \pi_G(1-(1-p_G)^2) + (1-\pi_G)(1-p_G^2) \end{aligned}$$

The first order condition is:

$$\pi_G(1 - p_G) = (1 - \pi_G)p_G$$

Since this condition is satisfied when the subject's stated forecast, p, is equal to her true belief, π, the quadratic scoring rule is incentive compatible or proper.

2. *Logarithmic Scoring Rule:* Although the logarithmic scoring rule does not satisfy the neutrality axiom, it does satisfy the other three and most significantly the incentive compatibility one. Hanson (2007) gives a summary of other potentially desirable properties of the logarithmic scoring rule. Under this rule, the subject's expected payoff given her true belief is:

$$\begin{aligned} E(PO) &= \pi_G(1 + 0.45 log p_G) + \pi_R(1 + 0.45 log p_R) \\ &= \pi_G(1 + 0.45 log p_G) + (1 - \pi_G)(1 + 0.45 log(1 - p_G)) \end{aligned}$$

The corresponding first order condition is:

$$\frac{\pi_G}{1 - p_G} = \frac{1 - \pi_G}{1 - p_G}$$

Again, the subject's stated forecast must be the same as her true belief to satisfy this condition and maximize her expected payoff. Therefore the logarithmic scoring is also proper.

3. *Linear Scoring Rule:* The linear scoring rule violates not only the neutrality axiom but more importantly the incentive compatibility axiom as well. The expected payoff is:

$$\begin{aligned} E(PO) &= \pi_G p_G + \pi_R p_R \\ &= \pi_G p_G + (1 - \pi_G)(1 - p_G) \end{aligned}$$

If the subject's true belief is the *Green* is more likely to be played, $\pi_G > 0.5$, then she maximizes her expected payoff under the linear rule by stating a forecast of 100% probability on event G. Similarly,if she believes that Red is more likely to be played, $\pi_G < 0.5$, then she maximizes her expected payoff by placing 100% probability on event R.

3 Experimental design and procedures

We conduced six sessions with a total of 48 subjects. The subjects were registered students at a US university,who were recruited by email solicitation. Sessions were conducted in a computer lab and all interaction was computerized. No subject participated in more than one session. The primary treatment variable was the scoring rule, either log, quadratic, or linear, with one third of the subjects in each treatment.

Each session had two parts. In the first part, subjects were randomly assigned to be either the row player or the column player in the 2x2 game in Table 1. Keeping the pairings fixed, they played the game repeatedly for 5 rounds. After round 5, they are assigned to the opposite role so that if they were a row player in the first five rounds, they are now a column player and vice versa. They are also randomly repaired with a different player and play the game repeatedly for 5 rounds with this new opponent. Their earnings for Part 1 was the sum of their earnings over all 10 rounds of play. The purpose of part 1 of the session was to familiarize the subjects with the strategic problem facing players of this game.

In part 2, subjects made "observer" forecasts about the sequence of choices of either the row or the column player in seven different pairs from the Nyarko-Schotter experiment. In each session, four of the subjects were assigned the task of forecasting behavior of row players (row forecasters) and four were assigned the task of forecasting the choices of column players (column forecasters) These roles were fixed throughout

part 2. For each of the seven pairs whose play they were asked to forecast, all subjects are told the actions chosen by both players in that pair in the first five rounds of their match. The matrix game and the list of actions chosen by the pair in the first five matches are displayed on every subject's computer screen throughout this part of the experiment. Row forecasters are then asked to report their probabilistic beliefs about the likelihood the row player in that pair chose red or green in round six, and column forecasters are asked to report their probabilistic beliefs about the likelihood the column player in that pair chose red or green in round six. This is implemented by requiring them to type in two positive integers, one for green and one for red, where the two numbers must add up to 100. All the column predictors simultaneously make forecasts in this manner about the actions of the same column player in round 6 of the same match from the NS experiment, and all the row forecasters simultaneously make forecasts in this manner about the actions of the same row player of the same match from the NS experiment.

After reporting these forecasts, all row forecasters are told the forecasts of all the other row forecasters, and all column forecasters are told the forecasts of all the other column forecasters. The subjects are then allowed to modify their original forecast. This is implemented by having them re-enter two positive integers, one for red and one for green, that sum up to 100. The revised forecasts of all the row predictors are shown to the row forecasters, and likewise for the column forecasters.

Subjects are then paid for their forecasts using a scoring rule that is described carefully in the instructions. Due to the possible incentive distortions that iterative elicitation may have on the proper scoring method, only one of the two forecasts a subject makes is randomly selected for payment.

After the second round of forecasts, the actual choices by the row and column players in round 6 of that NS pair are then reported back to the subjects, so they now

know the choices by both subjects in the first *six* rounds of the match. They are also told which of their two forecasts was randomly chosen for actual payoff, and the payoff is computed for them and appears on their screen. A history panel at the bottom of the client screen keeps track of all this information, and new information is appended to the history panel as the experiment proceeds. All subjects then proceed to make forecasts about round 7 of that NS pair, in the same manners as they made forecasts about round 6. Roles (row or column forecaster) stay fixed. As in round 6, subjects are allowed to revise their forecasts once, in light of the forecasts of other subjects in the same role. They continue in this way to make iterative forecasts for the play in rounds 8, 9, and 10 of that NS pair. This procedure was then repeated during the session for six other NS pairs. Thus, overall, subjects reported and revised forecasts about a total of 35 plays of the game by 7 different pairs. They were paid the sum of their earnings in each round.

Instructions were read aloud to the subjects. A copy of the instructions is available from the authors upon request. After the session ended, all subjects were paid one at a time in private. Total earnings, including a show-up bonus of \$10, ranged from \$17 to \$35. In the log and quadratic treatments, forecasts were constrained to be between 10 and 90. Because such a constraint is necessary in the log treatment to avoid bankruptcy problems, it was also included in the quadratic treatment for consistency. The constraint was not imposed for the linear scoring rule since the theoretical prediction for the linear scoring rule is at the boundary, i.e., forecast either 0 or 100.

4 Results

We analyze the results in two subsections. First, we describe the main aggregate features of the initial elicitation data, before subjects have had the chance to revise their forecasts in light of the forecasts of others. We compare the distribution of forecasts across treatments, across roles. We also compare our data with the distribution of

forecasts elicited from NS subjects in rounds 6-10 of that experiment and to the aggregate frequency of choices observed in their data.

Second, we analyze the accuracy of the forecasts. We use two benchmarks: uninformed forecasting (always forecasting 50/50) and rational expectations (forecasting the empirical average frequency in every round). We refer to 50/50 forecasts as uninformed because such a report is optimal for a forecaster whose prior is uniform on [0,1].

Third, we investigate questions about the iterative elicitation process. Does it lead to convergence of beliefs? Does the iterative process lead to better forecasts?

Finally, we try to answer the question of whether the distributions of beliefs we observe are biased in any systematic ways, compared to the forecasts of a fully rational model of belief elicitation with risk neutral subjects.

4.1 Individual forecasts: Comparison of scoring rules and comparison with NS

4.1.1 Distributions of initial forecasts

We first do a comparison of the first-round forecasts across the three treatments, averaged over all subjects in our experiment, broken down by role and scoring rule.

If the beliefs are the same for observers and players of the game, and if there are no distortions created by having a subject report beliefs and choose actions at the same time as in NS, then there should be no significant differences between the NS elicited beliefs and the beliefs elicited from observers using the quadratic scoring rule. Because the log scoring rule is also proper, there should be no differences between our log and quadratic treatments and there should be no differences between the log elicitations and the NS elicitations.

The first row of Table 2 describes the row players' elicited beliefs about the column player's likelihood of choosing green, and the second row describes the column player's

elicited beliefs about the row players' likelihood of choosing green. In this and subsequent tables, "column" refers to column moves or forecasts about column moves. "Row" refers to row moves or forecasts about row moves. The first three columns give the average forecast under our three scoring rule treatments. The fourth column is the average forecast in rounds 6-10 of NS experiment 1 (i.e. the same rounds our subjects were forecasting), and the final column gives the actual choice frequencies in those rounds.

Two observations are immediate, both concerning the accuracy of forecasts. First, the NS players and our own subjects systematically underestimate the probability column will choose green and overestimate the probability that row will choose green, but these differences are not significant . Second, this bias is less in all of our treatments and for both player roles than in the NS elicitation from the actual players. However, only the difference between the forecasts of the row player's actions under the linear scoring treatment and the NS forecasts is significant.

	Quad	**Log**	**Lin**	**NS Quad**	**Observed**
Column	45.7*	47.7*	45.5*	44.3	55.7
Row	48.8*	47.4*	43.4*+	53.0	42.9
N	560	560	560	140	140

Table 2. Average reported beliefs and actions by role and elicitation method.

* = less biased than NS forecasts. + = significantly less biased than NS forecasts (p=.05).

Another way to compare the forecasts of our observer subjects with the forecasts of the actual players of the game is to look at raw correlations between the two. Table 3 reports these raw correlations for row and column players separately and also pooling the two, using the average first round forecasts of each of our groups of four subjects,

matched with the forecasts of the corresponding NS subject. We find systematically positive correlations for our quadratic scoring rule treatment, but not for the log or linear scoring rules. To test for significance of these differences, we ran a Tobit regression of the mean of our elicited beliefs under quadratic treatment on the corresponding NS elicitations. The coefficient is significant at the 5% level for the quadratic treatment. In contrast, we cannot reject the hypothesis that our log and linear elicitations are uncorrelated with the elicited beliefs of the NS players. This suggests that the specific choice of elicitation method is an important factor since NS also used a quadratic scoring rule.

	Quad	**Log**	**Linear**
Row	0.094	−0.12	−0.023
Column	0.20	0.15	−0.052
Overall	0.17	−0.0081	−0.045

Table 3. Correlation between average Observer forecasts and matched NS forecasts.

	Quad	**Log**	**Linear**
Row	0.038(0.049)	−0.035(0.035)	−0.014(0.073)
Column	0.13(0.073)	0.037(0.029)	−0.056(0.12)
Overall	0.087 ∗ (0.043)	−0.0022(0.023)	−0.038(0.068)

Table 4. Coefficients of Tobit regressions of average Observer forecasts on NS forecasts. Standard errors in parenthesis. *=significantly non-zero (p=.05)

The results from Tables 3 and 4 show that the three scoring rules we use with observers clearly do lead to different measurements of beliefs. To explore this further, we examine the differences in *dispersion* across our three measures and look at how these dispersions compare with the NS elicitations. According to the theoretical results, we know that quadratic and log are both proper scoring rules, so we hypothesize no significant difference between the dispersion in beliefs for log and quadratic. In contrast,

the linear scoring rule is not proper; indeed, optimizing risk neutral subjects will report beliefs equal to either 0 or 1. We hypothesize the linear elicitation procedure will result in greater dispersion than the quad or log methods.

In addition, if the beliefs of the quadratic scoring rule were the same for observers and players of the game, and if there are no distortions created by having a subject report beliefs and choose actions at the same time as in NS, then there should be no difference between the player forecasts and the observer forecasts, at least under the two proper scoring rules. We hypothesize that there will be no differences in dispersion between NS forecasts and the observer forecasts using quadratic and log scoring rules.

To measure dispersion, we compute the absolute differences from 50 for each individual forecast. The average of these absolute differences across all forecasts in each treatment, broken down by row and column, are reported in Table 5, with the complete CDF of the differences displayed in Figure 1.

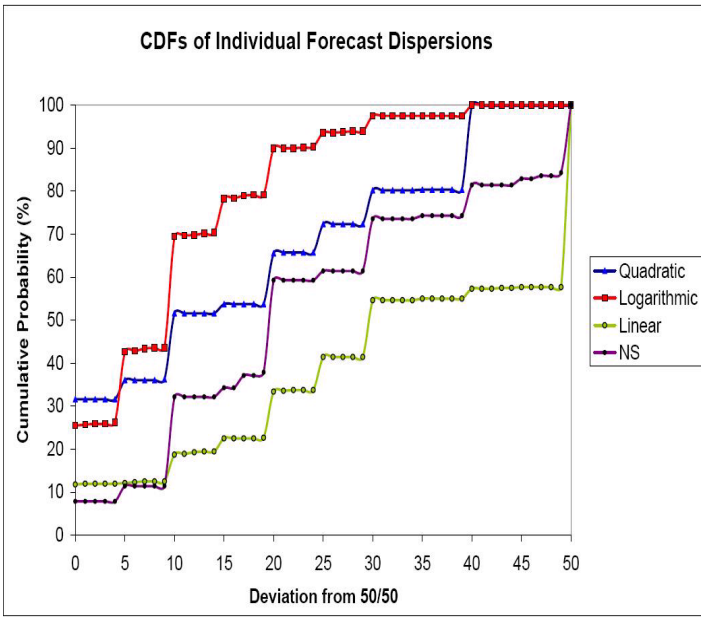

Figure 1. Individual forecast dispersions under the three scoring rules and NS.

The differences are striking. First, the NS player forecasts and linear forecasts exhibit more dispersion than the forecasts by observers with proper scoring rules. The differences are not only significant, but large in magnitude, with the NS dispersions more than double the log scoring rule dispersion and the 50% greater than the dispersion of observer quadratic elicited beliefs.

Second, the linear scoring rule leads to the greatest dispersion, with the comparison to log and quadratic significant for observers, as theory predicted. Overall, the linear forecasts are more dispersed than the NS forecasts and this difference is significant except for column players' behavior. Third, the observer forecasts using quadratic and log scoring rules are significantly different, with the dispersion under the quadratic scoring rule more than 50% more than under the log scoring rule.

	Quad	Log	Linear	NS
Column	13.53*(0.86)	11.45*(0.55)	26.68(1.07)	24.53(2.01)
Row	19.31*(0.92)	8.89*(0.57)	36.74*(1.01)	23.37(1.78)
Overall	16.42*(0.64)	10.17*(0.40)	31.71*(0.77)	23.95(1.34)

Table 5. Dispersion of beliefs, measured by individual average absolute difference from 50. Standard error in parenthesis.

*=significantly different from NS (p=.05)

Similar conclusions follow if one compares the dispersion of the *average* (across 4 subjects) first round forecast by our subjects rather than the individual dispersion (which adds an additional source of variance). The only difference is that the group-averaged dispersion is always less than the individual dispersions. Now, the ordering of NS and linear is reversed, due to group averaging, which eliminates the individual sources of variance of the absolute differences. However, this difference between NS and linear is not significant for the forecasts of row choices. These grouped dispersion measures are given in Table 6.

	Quad	Log	Linear	NS
Column	10.28*(0.90)	7.31*(0.62)	14.69*(1.30)	24.53(2.01)
Row	13.77*(1.24)	5.72*(0.51)	22.88(1.79)	23.37(1.78)
Overall	12.03*(0.78)	6.51*(0.41)	18.79*(1.16)	23.95(1.34)

Table 6. Dispersion of beliefs, measured by group-averaged absolute difference from 50. Standard error in parenthesis.

*=significantly different from NS (p=.05)

These differences can be clearly seen in Figure 1, which graphs the CDF of the *average* individual forecast absolute differences from 50 for the three observer treatments and NS, as well as in Figure 2, which graphs the CDF of the group average dispersions.

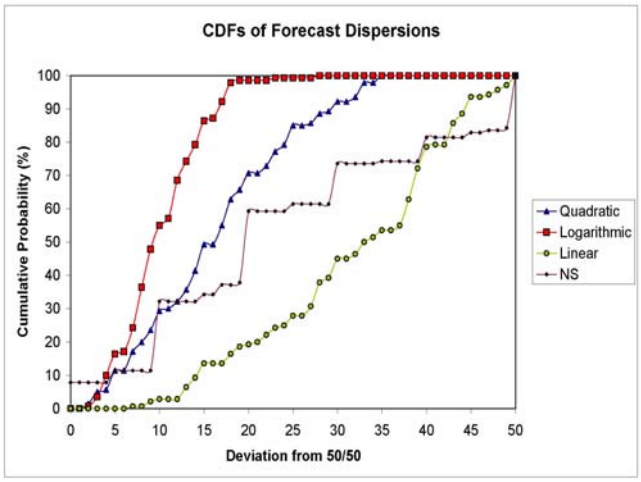

Figure 2. Group-averaged dispersions under the three scoring rules and NS.

We conclude that, in terms of extremeness, the elicited beliefs by players in the NS experiment closely resemble the beliefs that are elicited from trained observers using a linear scoring rule. We define extreme forecasts to be those that place 25% or less or 75% or more on the probability of Green being played. The NS and linear elicitation procedures resulted in the measurement of extreme forecasts 44.5% and 66.2% respectively compared to 34.3% for quadratic and 9.6% for log. Both procedures elicited significantly fewer fully-hedged (50/50) beliefs, even less frequently with NS than linear (7.9% vs. 11.8%). This is consistent with an interpretation of player-elicited beliefs in terms of "rationalization," whereby players report beliefs in a way that will be consistent with their actual choices.

4.1.2 Accuracy of reported beliefs: Do the subjects know anything?

In the actual NS experiment, and also in our experiment using trained observers, the subjects clearly think they know something. Over 92% of the time, they report "informed" beliefs, i.e., forecasts different from 50/50. We observe nearly the identical evidence that trained observers believe they have some information about behavior in our linear treatment, but considerably less so in the proper scoring rule treatments. See Table 7.

	Quad	**Log**	**Linear**	**NS**
Column	0.39	0.16	0.14	0.10
Row	0.25	0.35	0.09	0.06
Overall	0.32	0.26	0.12	0.08

Table 7. Fraction of forecasts exactly equal to 50/50.

It is then natural to ask whether this apparent confidence is justified. We find that for NS subjects, it is clearly unjustified. We document this more carefully in this section, but the bottom line is apparent from Table 2 in the previous section that shows forecasts of row and column actions to be systematically biased and on the wrong side of 50/50.

In contrast, we find evidence that the trained observers in our three treatments seem to have some forecasting ability. The first thing we look at is the raw correlation between forecasts and the choices they are forecasting. These are given in table 8. The overall correlations between forecasts and actions is highly significant for all three trained observer treatments. In contrast, one cannot reject the hypothesis that NS forecasts and actions are completely uncorrelated. Breaking this down by row and column forecasts, we find that in one case (column) they are positively correlated and in the other (row) they are negative correlated but not significantly. We also ran probit regressions of actual choice on observer forecasts by treatment and role and the patterns (sign and significance) in the coefficients match the correlations.

	Quad	Log	Linear	NS
Column	0.179∗	0.136∗	0.061	0.123∗
Row	0.121∗	0.027	0.157∗	−0.044
Overall	0.135∗	0.085∗	0.116∗	0.022

Table 8. Correlation between individual elicited forecast and actual choice in the experiment. *significantly non-zero (p=.05)

Table 9 compares the "hit rates" for the forecasts, where hit=1 if the green forecast is greater than 50 and green is chosen, or if the green forecast is less than 50 and red is chosen. Forecasts of 50/50 count as 1/2 a hit. Misses are scored as 0. Thus, completely uninformed forecasting will lead to a hit rate of .50.

	Quad	Log	Linear	NS
Column	0.561*(0.023)	0.527(0.027)	0.514(0.028)	0.521(0.023)
Row	0.559*(0.026)	0.539(0.024)	0.582*(0.028)	0.500(0.023)
Overall	0.560*(0.017)	0.533(0.018)	0.548*(0.20)	0.511(0.023)

Table 9. Hit rates by treatment. *significantly different from 0.50 (p=.05)

First, note that all the hit rates are .5 or higher, so subjects are at least doing no worse than uninformed forecasting! The NS hit rate is not significantly different from 0 for either column or row forecasts, or overall. Hit rates are generally higher for trained observers: in the quadratic treatment, both row and column forecasts are significantly informative at the 5% level, the overall and row forecasts are significant greater than .50 at the 5% level for the linear treatment and at the 10% level for the log treatment. So, by either measure, correlation or hit rates, the NS players are uniquely unjustified in believing they actually know something about how their opponent is likely to play.

A third way of assessing the accuracy of forecasts across scoring rules is to look at the mean squared deviation (MSD) of the forecast errors. For this we consider several baselines. The zero-information baseline is based on the MSD that would result by

always forecasting 50/50. This is equal to .25, and is a plausible upper bound on MSD for any *informed* forecaster who reports beliefs truthfully. The Nash equilibrium baseline is the MSD that would result, given the actual choice data, if one forecasted the Nash equilibrium mixed strategy every time. The NS forecast baseline is the MSD if an observer had simply used the NS player's forecast. The NS empirical baseline is the MSD if one always forecast the empirical round 6-10 choice frequencies of the player one whose behavior one is forecasting. It is a plausible MSD lower bound. Finally, we computed a baseline using the NS empirical choices using rounds 6-10 of all 14 pairs. These baselines are summarized in Table 10.

MSD	**Row**	**Column**	**Overall**
Uninformed	0.250	0.250	0.250
Nash	0.271	0.246	0.259
NS forecast	0.309	0.343	0.326
NS **empirical**	0.183	0.189	0.186
NS **empirical (all pairs)**	0.247	0.245	0.246

Table 10. MSD Baselines.

Table 11 gives the MSD scores for the quadratic, log, and linear scoring rule forecasts. None of the scores are below .25, suggesting that individuals forecasts on average have little more content than random guesses. The linear MSD scores are by far the worst and line up closely with the NS forecasts, which is consistent with other similarities between the two as noted above. However, this is partly due to the fact that the linear forecasts are almost certainly exaggerated (closer to 0 and 1 than true beliefs) as theory would predict. Since actual choice behavior is very close to 50/50, this exaggeration is "punished" by the MSD score in much the same way as a quadratic scoring rule would punish it. The log and quadratic scoring rules produce the lowest forecast error by this measure, with the log MSD lower than the quadratic MSD for both row and column forecasts. In fact, the log MSD is not significantly different from 0.25.

	Quad	Log	Linear	NS
Column	0.259	0.255	0.339*	0.309*
Row	0.280*	0.260	0.342*	0.343*
Overall	0.270*	0.257	0.341*	0.326*

Table 11. Average forecast errors measured by Mean Squared Deviation.

*=significantly different from 0.25 (p=.05)

Forecast Calibration We also use the calibration method (Seidenfeld 1985) to evaluate the accuracy of the subjects' forecasts. By Seidenfeld's definition, "a set of probabilistic predictions are *calibrated* if p percent of all predictions reported at probability p are true." A subject is perfectly calibrated in our experiment if for all the instances when she forecasted Green being played with 30% probability, Green is played 30% of the time, for all the time when she forecasted Green being played with 60% probability, Green is played 60% of the time, and so on.

We pool the forecasts across subjects to obtain a single aggregate calibration measure for each scoring rule. We also generate a comparable calibration measure for all subjects in the Nyarko-Schotter experiment, using their round 6-10 forecasts. We take every ten percentage points as one bin for the predicted probability of Green being played and calculate the corresponding Green action frequency for that bin. Each bin takes the lowest percentage as its value when we compare it to the frequencies for the calibration. For example, 10-19% is 10% probability that Green is played, 20-29% is 20% probability that Green is played, and so on. There is actually little distortion in this approach because most of the stated probabilities in each bin are at the lowest percentage.

Table 12 shows the corresponding frequency of Green action for each of the predicted probability bins under the different treatments in our experiment as well as in

the Nyarko-Schotter belief elicitation treatment. Bins that had less than five observations of play, 90-99% under logarithmic scoring rule and 10-19% in NS, were dropped. The numbers in this table generate the calibration curves in Figure 2. If the forecasts were perfectly calibrated, the line would have a slope of 1. Here we observe that the slope is significantly less than 1 for our three treatments and NS. The frequency of accurate forecasts, Green being played in this case, is less than the forecast probabilities suggesting that subjects were overconfident in their forecasts.
(ID=Insufficient Data: N<5)

Forecast	**Quadratic**	**Logarithmic**	**Linear**	**NS**
0	*ID*	*ID*	40.6	46.2
10	36.1	30.0	62.5	*ID*
20	47.7	48.1	40.7	46.2
30	38.5	41.2	53.3	41.7
40	37.5	49.7	52.4	58.8
50	52.9	48.2	57.4	50.0
60	61.7	56.7	55.2	52.6
70	58.2	60.6	42.5	45.5
80	64.7	83.3	52.6	50.0
90	51.3	*ID*	66.7	75.0
100	*ID*	*ID*	58.1	33.3

Table 12. Forecast of Green vs. frequency of Green.

ID=Insufficient Data (N<5)

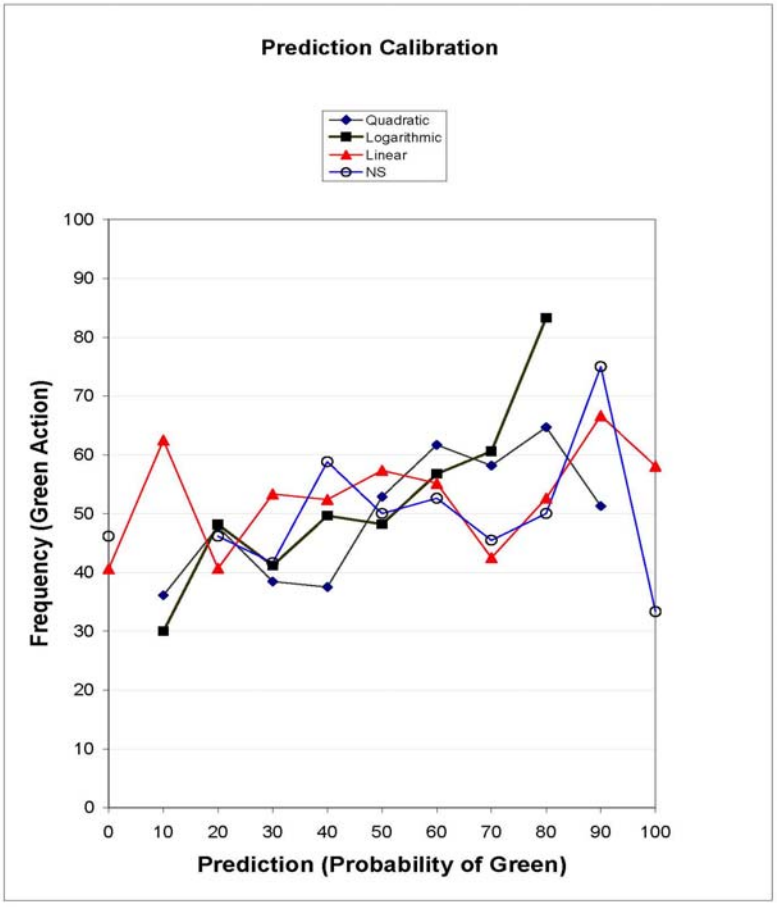

Figure 3. Forecast of Green vs frequency of Green.

We run OLS regressions of the action taken (100 for Green, 0 for Red) on the individual stated probability of Green being played and use the coefficient as a calibration measure. The coefficient would be 1 if the subjects are perfectly calibrated. As reported in Table 13, we find that the coefficients are all nonzero with statistical significance under all three of our treatments whereas we cannot reject that the

coefficient is zero for the NS experiment. Furthermore we find that the subjects are similarly calibrated under the quadratic and logarithmic scoring rule and are substantially less calibrated under the linear scoring rule.

Elicitation	**Constant**	**Green Prob.**	**N**	**R^2**
Quadratic	34.88*	0.30*	560	0.018
	(4.94)	(0.095)		
Logarithmic	34.50*	0.31*	560	0.0072
	(7.64)	(0.15)		
Linear	42.13*	0.16*	560	0.014
	(3.34)	(0.058)		
NS	47.64*	0.034	140	0.0004
	(8.40)	(0.15)		

Table 13. OLS regression of player actions on forecast probabilities. * = significantly different from 0 (p=.05)

4.2 Learning from others' forecasts

Our experiment had two key design features that allow us to look at questions of information aggregation. First, for each action decision to be forecast, we elicited forecasts from four trained observers, rather than just one. Second, there were two rounds of forecasts, and each forecaster was advised of the forecasts by the other forecasters before reporting a second round forecast. In this section, we address two specific questions about the effects of group feedback on forecasts and how the answers depend on the scoring rule.

1. Do subjects update their forecasts after learning the forecasts of the other forecasters? (*belief convergence*)

2. Are updated forecasts more accurate than initial forecasts? (*information aggregation*)

4.2.1 Belief Convergence

To address question 1, we first compute the percentage of the subjects change their forecast in the second round after being told the other forecasters' reports. The answer is yes, forecasters revise their reports in response to the reports of other forecasters.

	Quad	**Log**	**Linear**
Column	0.32	0.60	0.46
Row	0.41	0.54	0.38
Overall	0.37	0.57	0.42

Table 14. Frequency of revisions.

	Quad	**Log**	**Linear**
Column	3.31 (0.39)	5.20 (0.38)	10.73 (1.09)
Row	9.06 (0.96)	6.24 (0.53)	12.46 (1.44)
Overall	6.18 (0.53)	5.72 (0.33)	11.60 (0.90)

Table 15. Average revision (absolute value). Standard errors in parenthesis.

As further evidence, we look at the change in the variance of forecasts in the group, defined as the variance of second stage forecast - variance of first stage forecast. If the forecasts are closer together in the second round (negative change in variance), we take that to be evidence of belief convergence. Table 16 displays these change in variance by position and treatment. We do find that the within-group variance does decline from the first round to the second round for all three scoring rule treatments. Table 17 shows that while there were significantly more instances of decrease rather than increase in forecast variance in the quadratic and linear scoring rule treatments, the log treatment resulted in a close to even split before decrease and increase in forecast variance.

	Quad	**Log**	**Linear**
Column	-51.36*	-36.14*	-161.18*
Row	-77.24*	8.77	-23.04
Overall	-64.30*	-13.69*	-92.11*

Table 16. Changes in Variance. * = significantly different from 0 (p=.05)

	Quad	Log	Linear
Less Variance	0.51	0.49	0.61
No Change	0.24	0.04	0.10
More Variance	0.24	0.48	0.29

Table 17. Directional change in variance.

4.2.2 Information Aggregation

To address question 2, we compare the MSD of first round forecasts to the MSD of second round forecasts. Table 18 displays the average change in MSD (second round minus first round). The changes are negative in all cases except the linear forecasts of row choices, but the magnitudes are rather small and statistically insignificant. If we ask what percentage of the revised estimates are more accurate, we find that revised forecasts are more accurate more often than less accurate in all cases, but again, the differences are small. The exact numbers are given in Table 19.

	Quad	Log	Linear
Row	0.0042	-0.0011	0.0077
Column	-0.0123	-0.0012	-0.0168
Overall	-0.0040	-0.0012	-0.0045

Table 18. Average change in MSD.

	Quad	Log	Linear
More Accurate	0.44	0.49	0.54
No Change	0.24	0.03	0.09
Less Accurate	0.33	0.49	0.37

Table 19. Directional change in MSD.

5 Conclusions

The experiment reported here produced several findings on the elicitation of beliefs with scoring rules. First, the forecasts by our observers under all three scoring rules were more accurate than the forecasts of the NS players, in the sense that the average elicited forecast was closer to the true choice frequencies. Furthermore, there was a significant positive correlation between observer forecasts and the choice behavior in the game for

all three of our scoring rules, while there was no significant correlation between the players' forecasts and the actual play being forecasted. Second, the distribution of forecasts by NS players had more variance, in the sense of being more extreme, than the observer forecasts using either of the proper scoring rules. The distribution of forecast dispersions (differences from 50/50) most closely resembled the distribution of forecasts elicited from observers under the linear scoring rule with only slightly less dispersion. Third, the linear scoring rule does not always elicit extreme priors, but does so frequently and generally results in reported beliefs that are significantly more extreme than the two proper rules. The linear forecasts are also less well calibrated. Fourth, there is a significant difference between the elicited beliefs under quadratic and logarithmic scoring rules in spite of both being proper scoring rules. Forecasts elicited by the logarithmic scoring rule have significantly less dispersion. However, we did not observe differences in accuracy or calibration for the two rules. The relative accuracy of the two varied across our different accuracy measures. Fifth, we find significant evidence of both belief convergence and information aggregation.

A number of conclusions can be drawn from these results. We summarize our findings in terms of the answers they give to the four methodological questions and two substantive questions posed in the introduction of the paper.

1. *Can beliefs be reliably elicited from the players of a game, during the play of the game?* The first two findings described above raise serious doubts about the reliability of beliefs elicited from players who simultaneously have a stake both in the accuracy of their forecast and in the outcome itself, in this case an opponent's choice in a two person game. We also identified what one could call an *overconfidence bias* in player-forecasts, relative to the forecasts of experienced observers. A deeper exploration into the sources of this bias is an interesting topic for future research. A possible explanation is rationalization: players report

forecasts that reinforce their choice.

2. *Are forecasts elicited using proper scoring rules systematically different from those elicited by improper scoring rules?* Yes, as implied by the third finding listed above. Both proper scoring rules elicit forecasts from our observers that are significantly more accurate and better calibrated than those elicited under the linear scoring rule. The direction of the bias caused by linear forecasts is clear: the linear scoring rule elicits more extreme forecasts as predicted by standard theory. Nearly one-third of the forecasts elicited under proper scoring rules are fully hedged, compared to fewer than 10% under the linear rule. The large percentage of extreme forecasts under the linear rule, and the fact that these observations are scattered across an even higher percentage of our subjects, suggests that risk aversion is a relatively weak factor in our data.

3. *Do different proper scoring rules elicit similar forecasts?* Yes. The main difference between forecasts elicited under logarithmic and quadratic scoring rules was that the quadratic rule elicited more extreme beliefs than the logarithmic rule. The distribution of extremeness of forecasts under the quadratic rule stochastically dominates the distribution under the logarithmic rule. It is interesting that this did not result in either one eliciting more accurate or better calibrated forecasts on average than the other. Why we observe this difference is an open question. The procedures used were identical, except for the scoring rule, and it seems implausible that the difference is due to subject heterogeneity and sampling variation. Risk aversion is not a plausible explanation either. While risk aversion can distort reported forecasts, if subjects have constant relative risk aversion, there is virtually no difference in the theoretical distortion that would result under the two rules. Loss avoidance may be a possible explanation for the difference in

boundary forecasts, but cannot explain the stochastic dominance finding. Other possibilities, such as ambiguity aversion and other violations of expected utility theory are worth pursuing in future research, but are beyond the scope of this paper.

4. *Are elicited forecasts more accurate and/or better calibrated under some scoring rules than others*? Yes, but these differences are relatively small compared with the differences in how extreme the forecasts are. The accuracy of the rules is clearly ordered by the MSD measure. The log rule was the only rule that produced an MSD that was not significantly worse than always reporting an uninformative prior (50/50). The linear rule was clearly the worse under this measure, with an MSD very close to the MSD of the NS players' forecasts.

5. *Are individuals in a group able to update their beliefs in response to the forecasts of other members of the group?*

   We found significant frequency of forecast revisions from the first round to the second round in all three scoring rule treatments. The within group variance of second round forecasts is significantly less than the variance of first round forecasts. That is, forecasts are converging. Convergence is evidence that they are learning from each other.

6. *Are individual forecasts improved by group interaction?* Second round forecasts are more accurate than first round forecasts as measured by the MSD, although the magnitude of improvement is small and statistically insignificant. This suggests something beyond simply belief convergence, since it is theoretically possible for beliefs to be converging but become less accurate as precision and accuracy are two different things.

The choice of scoring rule to elicit probabilistic beliefs about subjective events can

make a big difference. The distribution of our elicited beliefs under the three scoring rules are significantly different from each other in important ways. Second, our results bolster support for evidence elsewhere that the elicitation of beliefs directly from players, simultaneously playing the game for which they are forecasting outcomes, is unreliable. In light of this, anomalies that have been cited in the literature about play being inconsistent with beliefs (e.g., Costa-Gomes and Weizsacker 2006) are not surprising. The evidence for unreliability is sufficiently convincing at this point, that a reasonable position might be that the use of such procedures provide data that is at best unreliable and at worst misleading. Our own view is more neutral, and one hopes that more reliable methods can be discovered. In the meantime, forecasts elicited directly from players should be interpreted cautiously with the expectation that they may be flawed in some of the ways identified here.

References

[1] Aumann, R. J. (1976): "Agreeing to Disagree" *Annals of Statistics* 4, 1236-1239.

[2] Brier, G. (1950): "Verification of Forecasts Expressed in Terms of Probability" *Monthly Weather Review,* 78, 1-3.

[3] Choi, S., D. Gale and S. Kariv (2005): "Behavioral Aspects of Learning in Social Networks: An Experimental Study" *Advances in Behavioral and Experimental Economics (Advances in Applied Microeconomics series)*, J. Morgan ed., JAI Press.

[4] Costa-Gomes, M. and G. Weizsacker (2006): "Stated Beliefs and Play in Normal Form Games" *Working Paper,* University of York, U.K.

[5] de Finetti, B. (1965): "Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item" *British Journal of Mathematical and Statistical Psychology,* 18, 87-123.

[6] Dominitz, J. and A. Hung (2004): "Homogeneous Actions and Heterogeneous Beliefs: Experimental Evidence on the Formation of Information Cascades" *Working Paper,* Carnegie Mellon University, U.S.A.

[7] Fountain, J. (2002): "Eliciting Beliefs from Risk Averse Forecasters Using a Log Scoring Rule" *Working Paper,* University of Canterbury, Christchurch, New Zealand.

[8] Geanakoplos, J. D. and H. M. Polemarchakis (1982): "We Can't Disagree Forever" *Journal of Economic Theory,* 28, 192-200.

[9] Glein, I. N. and J. B. Wallace Jr. (1974): "Probabilistically Answered Examinations: A Field Test" *The Accounting Review,* 49, 363-366.

[10] Goeree, J., C. Holt, and T. Palfrey (2003). "Risk Averse Behavior in Generalized Matching Pennies Games" *Games and Economic Behavior,* 45, 97-113.

[11] Good, I. J. (1952): "Rational Decisions" *Journal of the Royal Statistical Society B,* 14, 107-14.

[12] Hambleton, R. K., D. M. Roberts, and R. E. Traub (1970): "A Comparison of the Reliability and Validity of Two Methods for Assessing Partial Knowledge on a Multiple-Choice Test" *Journal of Educational Measurement,* 7, 75-82.

[13] Hanson, R. (2007): "Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation" *Journal of Prediction Markets*, forthcoming.

[14] Huck S. and G. Weizsacker (2002): "Do Players Correctly Estimate What Others Do? Evidence of Conservatism in Beliefs" *Journal of Economic Behavior and Organization*, 47, 71-85.

[15] Kadane, J. and R. Winkler. (1988), "Separating Probability Elicitation from Utilities" *Journal of the American Statistical Association*, 83, 357-63.

[16] Ledyard J., R. Hanson, and T. Ishikida (2005): "An Experimental Test of Combinatorial Information Markets" *Working Paper.*

[17] McKelvey, R. D., and T. Page (1986): "Common Knowledge, Consensus, and Aggregate Information" *Econometrica*, 54, 109-127.

[18] McKelvey, R. D., and T. Page (1990): "Public and Private Information: An Experimental Study of Information Pooling" *Econometrica*, 58, 1321-1339.

[19] McKelvey, R. D., T. Palfrey, and R. Weber (2000): "The Effects of Payoff Magnitude and Heterogeneity on Behavior in 2x2 Games with Unique Mixed Strategy Equilibria" *Journal of Economic Behavior and Organization,* 42, 523-48.

[20] Murphy, A. H. (1969): "On the 'Ranked Probability Score'" *Journal of Applied Meteorology,* 9, 360-364.

[21] Nelson, R. G. and D. A. Bessler (1989): "Subjective Probabilities and Scoring Rules: Experimental Evidence" *American Journal of Agricultural Economics,* 71, 363-369.

[22] Nielsen, L.T., A. Brandenburger, J. D. Geanakoplos, R. D. McKelvey, and T. Page (1990): "Common Knowledge of an Aggregate of Expectations" *Econometrica*, 58, 1235-1239.

[23] Nyarko, Y. and A. Schotter (2002): " An Experimental Study of Belief Learning Using Elicited Beliefs" *Econometrica*, 70, 971-1005.

[24] Ochs, J. (1995) "Games with Unique Mixed Strategy Equilibria: An Experimental Study" *Games and Economic Behavior*, 10, 202-217

[25] Offerman, T. and J. Sonnemans (1998): "Learning by Experience and Learning by Imitating Successful Others" *Journal of Economic Behavior and Organization,* 34, 559-575 .

[26] Offerman, T., J. Sonnemans, and A. Schram (1996): "Value Orientations, Expectations, and Voluntary Contributions in Public Goods" *Economic Journal*, 106, 817-45.

[27] Offerman, T., J. Sonnemans, G. van de Luilen, and P. Wakker (2006): "Correcting Proper Scoring Rules for Risk Attitudes" *Working Paper*, Erasmus University, Netherlands.

[28] Palfrey, T. and H. Rosenthal (1991): "Testing Game-Theoretic Models of Free Riding: New Evidence on Probability Bias and Learning" in *Laboratory Research in*

Political Economy (T. Palfrey, ed.), University of Michigan Press:Ann Arbor, 239-67.

[29] Phillips, L. D. and W. Edwards (1966): "Conservatism in a Simple Probability Inference Task" *Journal of Experimental Psychology*, 72, 346-354.

[30] Roberts, H. V. (1965): "Probabilistic Prediction" *Journal of the American Statistical Association,* 60, 50-62.

[31] Savage, L. J. (1971): "Elicitation of Personal Probabilities and Expectations" *Journal of the American Statistical Association*, 66, 783-801.

[32] Seidenfeld, T. (1985): "Calibration, Coherence, and Scoring Rules" *Philosophy of Science,* 52, 274-294.

[33] Selten, R. (1998): "Axiomatic Characterization of the Quadratic Scoring Rule" *Experimental Economics,* 1, 43-62

[34] Staël von Holstein, C.-A. S. (1970): "Measurement of Subjective Probability" *Acta Psychologica,* 34, 146-159.

[35] Staël von Holstein, C.-A. S. (1971): "An Experiment in Probabilistic Weather Forecasting" *Journal of Applied Meteorology,* 10, 635-645.

[36] Staël von Holstein, C.-A. S. (1972): "Probabilistic Forecasting: An Experiment Related to the Stock Market" *Organizational Behavior and Human Performance,* 8, 139-158.

[37] Winkler, R. L. (1968): "The Consensus of Subjective Probability Distributions" *Management Science*, 15, B61-B75.

[38] Winkler, R. L. (1969): "Scoring Rules and the Evaluation of Probability Assessors" *Journal of the American Statistical Association,* 64, 1073-1078.

[39] Winkler, R. L. (1971): "Probabilistic Prediction: Some Experimental Results" *Journal of the American Statistical Association,* 66, 675-685.

[40] Winkler, R. L. and A. H. Murphy (1968): "'Good' Probability Assessors" *Journal of Applied Meteorology,* 1, 751-758.