Journal of Economic Behavior & Organization 71 (2009) 98–109

Contents lists available at ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/jebo

On eliciting beliefs in strategic games

Thomas R. Palfrey*, Stephanie W. Wang

Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, United States

ARTICLE INFO

Article history:
Received 12 January 2008
Received in revised form 23 March 2009
Accepted 25 March 2009
Available online 10 April 2009

Keywords:
Scoring rules
Experiment
Game theory
Forecasting
Beliefs

ABSTRACT

Several recent studies in experimental economics have tried to measure beliefs of subjects engaged in strategic games with other subjects. Using data from one such study we conduct an experiment where our experienced subjects observe early rounds of strategy choices 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. We compare forecasts across the scoring rules and compare the forecasts of our trained observers to forecasts of the actual players in the original experiment. We find significant differences across scoring rules. The improper linear scoring rule produces forecasts closer to 0 and 1 than the proper rules, and these forecasts are poorly calibrated. The two proper scoring rules induce significantly different distributions of forecasts. We find that forecasts by observers under both proper scoring rules are significantly different from the forecasts of the actual players, in terms of accuracy, calibration, and the distribution of forecasts. We also find evidence for belief convergence among the observers.

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, there have been a number of recent attempts of direct measurement of probabilistic beliefs by experimental economists in the context of strategic games. Examples include Dominitz and Hung (2009) in the context of information cascades, Huck and Weizsäcker (2002) in the context of lottery choice experiments, McKelvey and Page (1990) for information aggregation, Dufwenberg and Gneezy (2000) in trust games, and Offerman et al. (1996) and Croson (2000) in voluntary contribution games. The results of those papers raise questions about 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 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

* Corresponding author.
E-mail address: trp@hss.caltech.edu (T.R. Palfrey).

0167-2681/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jebo.2009.03.025

2×2 asymmetric matching pennies game similar to the one originally studied by Ochs (1995) and more recently by McKelvey et al. (2000), Goeree et al. (2003), and Nyarko and Schotter (NS, 2002).

The first 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 fourth methodological question we address is whether forecasts elicited directly from the players of a game during the play of the game are systematically different from forecasts elicited from observers. Forecasts elicited from the players themselves could be confounded by a variety of effects including psychological factors such as rationalization or via distortion of incentives because they are also being paid according to their play in the game, which violates the "no-stakes" condition of Kadane and Winkler (1988). We address this question by comparing the accuracy and calibration of forecasts elicited from (experienced) observers to the forecasts elicited from 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 observers, while the sequence of player moves is played back to them in real time, using scoring rules to incentivize the forecasts. The NS players also made one-move-ahead forecasts under similar incentivized conditions while they were playing the game.

The substantive questions concern both convergence of subjective beliefs and information aggregation. First, are individuals in a group able to 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).

The experimental results produce several sets of findings. First, there are significant differences between the elicited beliefs under quadratic and logarithmic scoring rules in spite of both being proper scoring rules. Second, the improper linear scoring rule produces forecasts closer to 0 and 1 than the proper rules, as predicted by theory, and these forecasts are poorly calibrated. Third, the forecasts by our observers with both proper scoring rules were different from the forecasts of the NS players in terms of accuracy and calibration. Fourth, we find evidence for belief convergence among our observers.

1.1. Related literature

1.1.1. Scoring rules

Scoring rules, which yield payoffs as a function of a 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 others have been shown to be strictly proper. Savage (1971) specifies the general rule for generating the class of strictly proper scoring rules, and there have been numerous theoretical studies of desirable and undesirable properties of proper and improper scoring rules. Kadane and Winkler (1988) have studied the effect of risk aversion on forecasts under the quadratic scoring rule. Offerman et al. (in press) and Fountain (2002) propose procedures to adjust for non-neutral risk attitudes.

1.1.2. Previous experiments using scoring rules to elicit beliefs

The quadratic scoring rule is the most common one applied 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), outcomes of sporting competitions (Winkler, 1971), and game theory (see below). The logarithmic scoring rule has been applied to a much lesser extent in experiments on education testing (Hambleton et al., 1970; Glein and Wallace, 1974) and information aggregation (Ledyard et al., 2009).

A few articles in the psychology literature have studied belief elicitation with different scoring rules, but none has conducted a comprehensive comparison of elicitations from the logarithmic, quadratic, and linear scoring rules, none have looked at the use of scoring rules for belief elicitation in the context of strategic choices in games, and none have compared player and observer forecasts.

Studies in experimental economics that have tried to use scoring rules to elicit subjective beliefs about action choices in a strategic game have produced mixed results. In the context of two-person matrix games, extreme reported beliefs

are observed with surprising frequency (Nyarko and Schotter, 2002). Because the "true" frequencies of target states are generally between 0.35 and 0.65 in these studies, this suggests bias 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, 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 should have a smoother trajectory than the observed forecasts. There is also evidence from two person laboratory games that the process by which subjects decide on a forecast is qualitatively different from the process they use to make a decision, which can sometimes result in forecasts that are inconsistent with choice behavior (Costa-Gomes and Weizsäcker, 2008).

In contrast, Dominitz and Hung (2009), 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 possible distortions in the elicitation procedure. The task was different from our task of one-step-ahead forecasts of choices in a repeated game since their subjects were repeatedly forecasting a static target (the state of the world) rather than a stochastically moving target. Offerman et al. (1996) elicited subjective player forecasts 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 exhibit an optimism bias.

There is very little evidence about the similarities and differences between forecasts elicited from observers and forecasts elicited from players themselves, and what evidence exists is mixed. Huck and Weizsäcker (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 does not 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. These two findings are also at odds with findings reported in Offerman et al. (1996, p. 828), where observers submitted forecasts that were more extreme than those submitted by the players themselves.

Rutström and Wilcox (2008) examine the effect of forecast elicitation on the choices of the forecaster and find significant effects. This provides evidence of a somewhat different confound between player forecasts. They also find that player forecasts are less accurate than empirically predicted beliefs based on simple statistical learning models.

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.

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; 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 an unincentivized 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

Table 1 displays the simple matrix game that was used in the Nyarko-Schotter experiment and in ours as well. This is a constant sum game with a unique Nash equilibrium in mixed strategies. 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 forecaster maximizes her expected monetary payoff by revealing her true belief. We next describe 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.

Table 1
Matrix game payoffs.

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

2.2.1. Preliminaries

Let $i = 1, 2, \ldots, n$ denote the n possible events and let $p = (p_1, p_2, \ldots, p_n)$ be the forecaster's stated forecast, where p_i is the stated probability of event i. Define the scoring rule $S = \{S_1, S_2, \ldots, S_n\}$ as a collection of scoring functions where $S_i(p)$ specifies the score when event i is realized as a function of the forecast, p. Let $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ be the subject's true belief where π_i is the probability of event i.

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 $\alpha, \beta > 0$ and I_k is an indicator function that takes the value 1 if the realized event is event k and 0 otherwise. The quadratic rule scores the inaccuracy of the forecast as a constant minus the sum of the squared deviations. In our belief elicitation experiment, there are two possible events: the event that the player being observed chooses Green, which we denote as G, or Red, R. We denote the two forecasts by p_G and p_R, respectively, where $p_G + p_R = 1$. Following Nyarko and Schotter (2002), we pay our subjects in the quadratic treatment an amount in dollars that is proportional to their score, using parameters $\alpha = 1$ and $\beta = 0.5$. The score is therefore

$$\begin{array}{ll} S_G = 1 - p_R^2 & \text{if} \quad G \text{ is chosen} \\ S_R = 1 - p_G^2 & \text{if} \quad R \text{ is chosen.} \end{array}$$

It is a straightforward exercise to prove that the quadratic rule is proper: a forecaster with true beliefs π maximizes expected score (expected payoff) by reporting $p = \pi$.

2. *Logarithmic scoring rule*:

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

where $\alpha, \beta > 0$.

The logarithmic rule, which is also proper, equals a constant less a penalty proportional to the natural log of the forecast of the realized event (a negative number since $0 \leq p_i \leq 1$). The lower the forecast of the realized event, the greater is the penalty. We set $\alpha = 1$ and $\beta = 0.45$. The score if event i occurs in the logarithmic treatment is

$$S_i(p) = 1 + 0.45 \sum_{k=1}^{n} I_k \ln(p_k).$$

3. *Linear scoring rule*:

$$S_i(p) = \alpha + \beta \sum_{k=1}^{n} I_k p_k \tag{3}$$

where $\beta > 0$.

We use $\alpha = 0$ and $\beta = 1$ in our experiment, so the linear score is simply the probability forecast for the realized event. The linear scoring rule is not proper. A forecaster with true beliefs π maximizes expected linear score by placing maximum weight on the most likely event. If the forecaster believes the two events are equally likely, then any forecast is optimal.

3. Experimental design and procedures

We conducted six sessions with a total of 48 subjects. Subjects were registered students at Princeton University and were recruited by email solicitation. Sessions were conducted at the Princeton Laboratory for Experimental Social Science, and all

interaction was computerized. Each subject participated in exactly one session, with eight subjects per 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. Instructions were read aloud to the subjects.[1] In the first part, subjects were randomly assigned to be either the row player or the column player in the 2×2 game in Table 1. Keeping the pairings fixed, they played the game repeatedly for five 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 re-paired with a different player and play the game repeatedly for five rounds with this new opponent. Their earnings for Part 1 was the sum of their earnings over all 10 rounds of play. The sole purpose of part 1 of the session was to give subjects experience with the game.

In part 2, subjects did not play the game but instead made "observer" forecasts about the sequence of choices of either the row or the column player in seven different pairs from the Nyarko–Schotter (NS) experiment. In each session, four subjects (row forecasters) were assigned the task of sequentially forecasting choices of NS row players and the other four subjects (column forecasters) were assigned the task of forecasting the choices of NS column players. These roles were fixed throughout part 2. The scoring rule (quadratic, log, or linear) was fixed throughout the session and was explained carefully to the subjects.

We then played back the data sequentially to the subjects in the following way. First, for one particular NS pair, all eight observers are told the actions chosen by the two players of that NS pair in the first five rounds of that match. The list of actions chosen by that NS pair in the first five matches is displayed on every subject's computer screen. Each row forecaster is then asked to report a forecast about the likelihood the row player in that pair chose red or green in round 6, and column forecasters are asked to report a forecast about the likelihood the column player in that pair chose red or green in round 6. This is implemented by requiring each subject to type in two integers, one for green and one for red, where the two numbers must add up to 100.[2] All the column predictors simultaneously and independently make forecasts in this manner about the actions of the one column player in round 6 of that NS pair, and all the row forecasters simultaneously and independently make forecasts in this manner about the actions of the one row player of the same NS pair.

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. We then elicit a second forecast from each subject by the same method. This second forecast can be the same or different from the first forecast.

After the revised forecasts have been made, 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. For each subject, one of their two forecasts was randomly chosen for actual dollar payoff.

Subjects then proceed to make forecasts about round 7 of that NS pair, in the same manner as they made forecasts about round 6. Roles (row or column forecaster) stay fixed. They continue in this way to make iterative forecasts for the play in rounds 8, 9, and 10 of that NS pair, receiving feedback after each forecast. This procedure was then repeated (sequentially) during the session so that the eight subjects observed a total of seven NS pairs. Thus, overall, subjects reported and revised forecasts sequentially for a total of 35 plays of the game by seven different pairs. They were paid the sum of their dollar scores in all 35 rounds. Total earnings ranged from \$17 to \$35.

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 and 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 more accurate forecasts?

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

Table 2 compares the average forecasts and the actual choice frequencies, broken down by scoring rule and by role (row or column).[3] 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

[1] A sample copy of the instructions is in Appendix A.

[2] Because the log scoring rule gives infinitely negative payoffs at the boundary (0 or 100), forecasts for that scoring rule were constrained to be in the interval [10, 90]. For consistency and to allow comparisons across the different scoring rules, the same constraint was imposed with the other scoring rules.

[3] The analysis in this section considers only the *first* elicited forecast of subjects. These beliefs are made before they know the forecasts of the other members of their group. We analyze the revised forecasts in the next section, where we address questions of convergence of beliefs and information aggregation.

Table 2
Average forecasts compared to observed choices. Entries are % Green.

	Quad	Log	Lin	NS quad	Observed
Column	45.7[a]	47.7[a]	39.8	44.3	55.7
Row	48.8[a]	47.4[a]	51.7[a]	53.0	42.9
N	560	560	560	140	140

[a] Less biased than NS forecasts.

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

	Quad	Log	Linear
Correlation	0.17	−0.0081	0.072
Tobit coefficient	0.087* (0.043)	−0.0022 (0.023)	0.037 (0.043)

Standard errors in parenthesis.
* Significant $p < 0.05$.

Table 4
Forecast extremeness: average absolute difference from 50.

	Quad	Log	Linear
Extremeness	16.42 (2.27)	10.17 (1.28)	19.38 (2.61)

Standard errors in parenthesis.

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.

Three results are illustrated by this table. First, the NS players and our own subjects systematically underestimate the probability that 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 both observer treatments with a proper scoring rule, and for both player roles, compared to the NS elicitation from the actual players.[4] Third, for observers, the bias is less with the proper scoring rules than with the linear scoring rule.

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. The first row of Table 3 reports these raw correlations 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 large positive correlations for our quadratic scoring rule treatment, less so for the linear rule, and actually negative for the log rule. To test for significance of these differences, we ran a Tobit regression of forecasts under each observer treatment on the corresponding NS elicitations. Rather than using individual observations, all forecasts from members of the same four person group were combined into one average forecast. This reduces the sample size by a factor of 4 relative to using all individual observations as a hedge against inflated significance levels. The coefficients and standard errors are reported in the second row of Table 3. The coefficient is significant at the 5% level only for the quadratic treatment. Based on the Tobit estimates, we cannot reject the hypothesis that our log and linear elicitations are uncorrelated with the elicited beliefs of the NS players.

The results from Tables 2 and 3 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 *extremeness* of elicited beliefs across our three measures. To measure extremeness, we compute the absolute differences from 50 for each individual forecast. 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 forecasts 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.

The average extremeness across all observer forecasts for each of the three scoring rules is reported in Table 4, with the complete CDF of the differences displayed in Fig. 1.[5]

The differences are striking. First, the linear scoring rule leads to the greatest dispersion among the observers, with the comparison to log and quadratic as predicted.[6] Second, our subjects' forecasts using quadratic and log scoring rules are significantly different from each other, with the dispersion under the quadratic scoring rule 60% more than under the log scoring rule.

[4] Because the observer forecasts were limited to the range of 10–90 and the NS forecasts were not, this could be the direct result of this truncation. As a check, we have replicated the analysis of NS forecasts in Table 2 by recoding those forecasts that are more extreme than 10 and 90, as 10 and 90, respectively. The results are the same. A similar replication was done for Tables 4 and 5 as well.

[5] Because the NS elicitation procedure allows more extreme forecasts, that distribution of extremeness is not directly comparable to our observers.

[6] This is consistent with Nelson and Bessler (1989) who found that a linear scoring rule generated more extreme forecasts than a quadratic scoring rule.

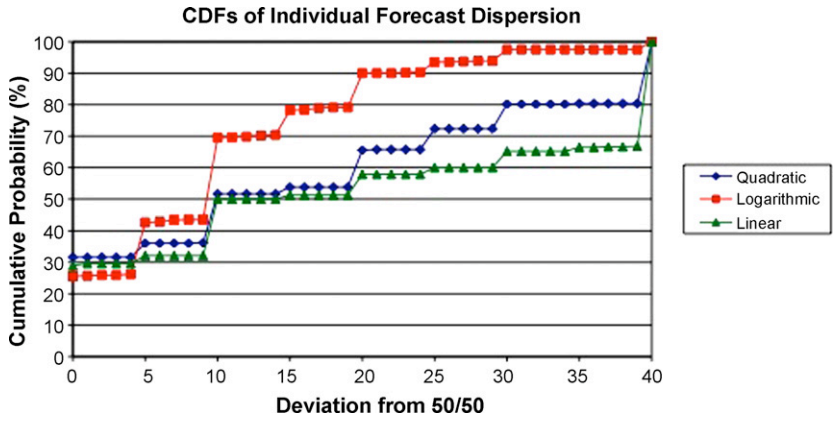

Fig. 1. Individual forecast dispersions for the three scoring rules.

Table 5
Correlation between individual elicited forecast and actual choice in the experiment.

	Quad	Log	Linear	NS
Correlation	0.135	0.085	−0.085	0.022
Calibration regression coefficient	0.30* (0.15)	0.31* (0.22)	−0.17 (0.11)	0.034 (0.15)
Calibration regression constant	34.9+ (6.5)	34.5+ (10.5)	57.0 (5.9)	47.6 (8.4)

Clustered standard errors in parenthesis.
* Coefficient significantly greater than 0.
+ Constant term significantly less than 50.

Three other features of the distribution of extremeness are worth noting. The first is stochastic dominance. The distribution of extremeness varies across the three scoring rules. Comparing the two proper scoring rules, the distribution of the quadratic rule stochastically dominates the logarithmic rule, except for an insignificant difference at 0, and both proper scoring rules for observer forecasts are stochastically dominated by the improper rule (again with an insignificant difference at 0). The second feature about the distributions of interest is the frequency of boundary forecasts (i.e., forecasts of 10% or 90%). The linear elicitation procedures resulted in the most boundary forecasts (33.2%), with the proper scoring rules having significantly fewer (19.6% for quadratic and 2.5% for logarithmic). The third feature is the percentage of 50/50 forecasts. Our subjects did this approximately 30% of the time, with insignificant variation across the three scoring rules.

4.1.1. Accuracy of reported beliefs: do the subjects know anything?

In the vast majority of cases, subjects report beliefs different from 50/50, reflecting confidence that they can predict behavior better than pure chance. It is then natural to ask whether the apparent confidence of the players is justified. The evidence suggests it is not. We document this in detail below, but the bottom line is apparent from Table 2 in the previous section that shows NS forecasts of row and column actions to be systematically biased and on the wrong side of 50/50. Moreover, because the choice behavior aggregate frequencies hover around 50% green, extreme forecasts seem to be harder to defend as "rational", compared with fully hedged forecasts.

In contrast, we find evidence that the trained observers *with proper scoring rules* seem to have some forecasting ability. First we look at the raw correlation between forecasts and the choices they are forecasting. These are given in the first row of Table 5. Second, we ask how well calibrated the forecasters are (Seidenfeld, 1985). By Seidenfeld's definition (p. 275), "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. Table 6 shows the frequency of Green choice for each forecast (pooled into bins 0–10, 11–20, etc.), with the number of observations in parenthesis. It is clear from Table 6 that the NS forecasts and the ones under the linear scoring rule are badly calibrated.

In order to make statements about the statistical significance of calibration, we ran calibration regressions of the action taken (100 for Green, 0 for Red) on the (first round) forecasts of Green being played. The coefficient on the action choice would be 1 and the intercept 0 if the subjects are perfectly calibrated. The coefficient would be 0 and the intercept 50 if the subjects' forecasts are perfectly uninformed. As reported in the second and third rows of Table 5, we find that the coefficients are significantly greater than 0 (quad $p < 0.05$; log $p < 0.10$) and the intercept is significantly less than 50 (quad $p < 0.05$; log $p < 0.10$) for the observer treatments with proper scoring rules. In contrast, the coefficients are not significantly greater than 0, and the constant terms are not significantly less than 50 for either the NS players or the observers using an improper scoring rule.

Table 6
Calibration: observed percent green choice by forecast.

Forecast bin	Quadratic	Logarithmic	Linear	NS
0–10	36.1 (71)	30.0 (10)	53.1 (113)	50.0 (16)
11–20	42.9 (28)	35.7 (14)	48 (25)	41.7 (12)
21–30	44.9 (49)	46.0 (50)	45.5 (22)	45.5 (22)
31–40	41.2 (51)	47.2 (125)	57.1 (77)	55.6 (18)
41–50	52.4 (189)	47.9 (192)	48.3 (174)	46.2 (13)
51–60	53.6 (56)	56.3 (103)	61.8 (34)	47.1 (17)
61–70	51.3 (39)	48.9 (47)	38.1 (21)	50 (14)
71–80	71.1 (38)	86.7 (15)	47.4 (19)	71.4 (7)
81–89	(0)	(0)	50 (2)	0 (1)
90–100	51.3 (39)	0 (4)	37.0 (73)	50 (20)

4.1.2. Constrained forecasts

In order to allow direct comparisons between the logarithmic scoring rules and the other two scoring rules, linear and quadratic, our experimental design required forecasts between 10 and 90 (inclusive of the bounds). In contrast, the NS players were permitted to submit any forecast between 0 and 100. Thus, it is useful to ask exactly how the forecast constraint might affect the distribution and accuracy of forecasts. Theoretically, the difference should be innocuous, in the sense that if a subject's unconstrained optimal forecast lies between 0 and 10 or between 90 and 100, the optimal constrained forecast would be 10 or 90 respectively. This is the only effect of the constraint in theory. That is, if a subject's unconstrained optimal forecast lies between 10 and 90, the optimal constrained forecast is equal to the optimal unconstrained forecast, and therefore there should be no distortion to these forecasts.

Thus there are two implications of this truncation. First, there will be an effect on the distribution of extremeness of forecasts. Specifically, the average extremeness will be higher if forecasts are unconstrained. Second, there can be a small effect on the calibration of subject forecasts because we observe their truncated forecasts rather than their optimal unconstrained forecasts. For example, an unconstrained and perfectly calibrated subject should have a calibration regression coefficient of 1 and an intercept of 0 in a very large sample. However, if such a subject is constrained to report forecasts between 10 and 90, the expected estimated coefficient would be slightly less than 1 and a significantly positive intercept, even in a large sample.

There are also possible psychological or framing effects that one could speculate about. Therefore, we conducted an additional comparison treatment with the quadratic scoring rule, where our observer subjects were allowed to make any forecast between 0 and 100.[7] This allows us to see what, if any, differences between the NS player forecasts and our observer forecasts might be attributable to truncation. We find that the distribution of forecasts by observers using the quadratic scoring rule without truncation are more extreme than with truncation, as theoretically predicted. The average extremeness increases from 16.42 with truncation to 23.76 without truncation, which is nearly identical to the average extremeness in the NS data (23.95). The CDF of extremeness of unconstrained observer forecasts under the quadratic rule is similar to the NS data based on player forecasts. In other respects (correlation, calibration, etc.), truncation has no significant effect.

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 others' forecasts? (*belief convergence*)
2. Are updated forecasts more accurate? (*information aggregation*)

4.2.1. Belief convergence

To address question 1, we first compute the frequency that subjects change their forecast in the second round after being told the other forecasters' reports and the average revision. The findings are reported in Table 7. The answer is yes; forecasters revise their reports in response to the reports of other forecasters. The frequency of revision ranges approximately 1/3 to 1/2, and the average absolute change is significantly positive for all three scoring rules ($p < 0.001$).[8]

The finding that forecasts change from the first to the second round leaves open the question of whether forecasters adjust in the direction of the other forecasts.[9] In principle, these changes could simply reflect random fluctuations, completely unrelated to the forecasts of the other members of the group. To explore this more carefully, we ask whether

[7] These sessions followed the same procedures as the other observer sessions.
[8] As in the calibration analysis, our hypothesis is signed so we conduct one-tailed tests.
[9] We are grateful to a referee for suggesting we dig deeper into the question of forecast adjustment.

Table 7
Frequency of and average revisions.

	Quad	Log	Linear
Frequency	0.37	0.57	0.32
Average	6.18* (1.64)	5.72* (1.14)	8.73* (2.28)

* Significantly positive.

Table 8
Directional change in variance.

	Quad	Log	Linear
% Less variance	51	49	44
% No change	24	4	29
% More variance	24	48	28
Average change	−64.30*	−13.69*	−35.57*

* Significantly different from 0 ($p < 0.05$).

an individual's second round forecasts are closer to the first round forecasts of the other members of the group than the individual's first round forecasts. Thus, for each individual i and each period t, we compute the following difference:

$$\Delta_i^t = |f_{i1}^t - F_{-i1}^t| - |f_{i2}^t - F_{-i1}^t|$$

where

$$F_{-i1}^t = \frac{1}{3}\sum_{j \neq i} f_{j1}^t.$$

The sum is taken over the three other members of i's group for that period. We then test the hypothesis that $\Delta_i^t > 0$. The difference is positive for all three scoring rules. This movement to the mean is significant for the pooled sample and is also significant for the quad treatment separately ($p = 0.02$). It is not significant for the log and linear subsamples.

A third hypothesis about the source of belief convergence is that forecasters adjust more if their own initial forecast was an outlier. That is, we hypothesize that $\Delta f_i^t = |f_{i1}^t - f_{i2}^t|$ is increasing in $|f_{i1}^t - F_{-i1}^t|$. The regression coefficients are significantly positive for all scoring rules ($p < 0.01$ for pooled sample and for quad and log subsamples; $p = 0.01$ for linear subsample).[10] We can also ask whether forecasts adjust more depending on how "wrong" the initial forecast was, where wrongness is measured by the distance between f_{i1}^t and the true state (either 0 or 1). Regressing Δf_i^t on wrongness of beliefs yields positive coefficients for all three scoring rules (significant at $p < 0.10$ for pooled sample and quad subsample). Finally, we can compare the forecast adjustment of subjects depending on how much variance they exhibit in first round forecasts. That is, we hypothesize that subjects whose first round forecasts are highly volatile across periods will adjust more than subjects with relatively little variation. Measuring volatility by the standard deviation across periods of a subject's first round estimates, we regress Δf_i^t on volatility and test the hypothesis that the coefficient is negative. The estimated coefficients are negative and highly significant for all three scoring rules (quad $p = 0.02$; log $p < 0.001$; linear $p < 0.01$). We also ran a multivariate regression of Δf_i^t on all three variables (outlier, wrongness, and volatility). All the predicted signs are correct for all three scoring rule subsamples, and all except two (wrongness coefficients for quad ($p = 0.06$) and linear subsamples) are significant at $p < 0.05$.

As a final piece of evidence about belief convergence, we look at the change in the variance of forecasts in the group, defined as the variance of second stage forecast minus 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. The first three rows of Table 8 display the percentage of times the change was negative, zero, or positive, by scoring rule. We find that the within-group variance declines from the first round to the second round about half the time for all three scoring rule treatments and declines more frequently than it increases. The last row of Table 8 gives the average change in variance for each scoring rule. For all scoring rules, the average change is significantly different from zero ($p < 0.05$) and negative.

4.2.2. Information aggregation

To address question 2, we look at the difference between the mean squared deviation (MSD) of initial forecasts and actions versus revised forecasts and actions. The first three rows of Table 9 display the percentage of times the change (revised minus

[10] In these tests, correlations due to individual differences are dealt with by using clustered standard errors. The measures we have constructed, such as $|f_{i1}^t - F_{-i1}^t|$, may result in some additional correlation across observations that is not accounted for.

Table 9
Directional change in MSD.

	Quad	Log	Linear
% More accurate	19	31	18
% No change	63	43	68
% Less accurate	18	26	15
Average change	−0.0040	−0.0012	−0.012
Standard error	0.0063	0.0042	0.0089

initial) was positive, zero, or negative, respectively, by scoring rule. We find that revised forecasts are more accurate by this measure than initial forecasts, but the differences are not large. The last row of Table 9 shows the average change in mean square deviation of forecasts from action (revised minus initial). The changes are negative in all cases, but the magnitudes are small and not statistically significant.

5. Conclusions

The experiment reported here produced several findings on the elicitation of beliefs with scoring rules. First, forecasts elicited from observers under the proper scoring rules were significantly more accurate and better calibrated than those elicited from players and from observers using an improper scoring rule. Second, there is a significant difference between distribution of 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. Third, the linear scoring rule elicits forecasts that are significantly more extreme than forecasts elicited by the two proper rules. Fourth, there was a significant positive correlation between observer forecasts and the choice behavior in the game for both proper scoring rules, while there was no significant correlation between the players' forecasts and the actual play being forecasted; the correlation was actually negative for the improper scoring rule (Table 5). Fifth, the forecasts by our observers under both proper scoring rules were less biased than the forecasts of the NS players, in the sense that the average elicited forecast was closer to the true choice frequencies (Table 2). Sixth, the distribution of forecasts by NS players was more extreme than the observer forecasts using either of the proper scoring rules. The average NS player forecast deviations (differences from 50/50) were not significantly different from forecasts elicited from observers under the linear scoring rule. Seventh, we find significant evidence for belief convergence but only marginal evidence for information aggregation.

The subset of findings about differences between observer forecasts and player forecasts are tempered by the fact that the experiments used slightly different instructions (by necessity, since our subjects were observers), slightly different subject pools (NYU students vs. Princeton students), and limited forecasts to be greater than 0.09 and less than 0.91 (because of the log scoring rule). The latter design variation directly affects the extremeness of elicited forecast by truncation. When our observers are unconstrained and use a quadratic scoring rule, their forecasts are as extreme as the forecasts of the NS players.

A number of conclusions can be drawn from these findings. 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. *Are forecasts elicited using proper scoring rules systematically different from those elicited by improper scoring rules?* Yes, as implied by the findings 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. One source of the bias caused by linear forecasts is that it elicits more extreme forecasts, as predicted by standard theory.
2. *Do different proper scoring rules elicit similar forecasts?* No. 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.
3. *Are elicited forecasts more accurate and/or better calibrated under some scoring rules than others?* Yes. Forecasts from proper scoring rules are more accurate and better calibrated than forecasts from improper scoring rules.
4. *Can beliefs be reliably elicited from the players of a game during the play of the game?* We find some evidence 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. The player forecasts are more biased than the

observer forecasts. The player forecasts are uncorrelated with the choice behavior they are forecasting, in contrast to the forecasts of observers under proper scoring rules.

5. *Do individuals in a group update their beliefs in response to the forecasts of other members of the group?* Yes. We found significant forecast revisions in all three scoring rule treatments. The within group variance of revised forecasts is significantly less than the variance of initial forecasts. We infer from this that beliefs are converging.
6. *Are individual forecasts improved by group interaction?* Revised forecasts are more accurate than initial forecasts, as measured by the MSD, but the magnitude of improvement is small and statistically insignificant. We conjecture that this is due to the fact that subjects were predicting outcomes of random variables whose underlying distribution was close to uniform (i.e., choices of player in a mixed strategy game with equilibrium at [0.6, 0.4]). It would be interesting to compare scoring rules in a game theoretic setting with more extreme probabilities, such as the games studied by Ochs (1995), Goeree et al. (2003), or Rutström and Wilcox (2008).

The choice of scoring rule to elicit probabilistic beliefs about subjective events can make a difference. The distribution of our elicited beliefs under the three scoring rules are significantly different from each other. Also, our findings add to evidence from other studies that the elicitation of beliefs directly from players simultaneously playing the game for which they are forecasting outcomes may be problematic and should be interpreted cautiously until these issues are better understood.

Acknowledgements

We gratefully acknowledge the financial support of the National Science Foundation (SES-0617820), The Gordon and Betty Moore Foundation, the Princeton Laboratory for Experimental Social Science, and the Social Science Experimental Laboratory at Caltech. We are grateful for detailed comments and suggestions from an editor, two referees, Juan Carrillo, audience members at the 2007 meeting of the Public Choice Society in Amsterdam, and participants at the 2007 conference at the University of Exeter on Risk, Forecast, and Decision.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2009.03.025.

References

Aumann, R.J., 1976. Agreeing to disagree. Annals of Statistics 4, 1236–1239.
Brier, G., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1–3.
Costa-Gomes, M., Weizsäcker, G., 2008. Stated beliefs and play in normal form games. Review of Economic Studies 75, 729–762.
Croson, R.T.A., 2000. Thinking like a game theorist: factors affecting the frequency of equilibrium play. Journal of Economic Behavior and Organization 41, 299–314.
Dominitz, J., Hung, A., 2009. Empirical models of discrete choice and belief updating in observational learning experiments. Journal of Economic Behavior and Organization 69, 94–109.
Dufwenberg, M., Gneezy, U., 2000. Measuring beliefs in an experimental lost wallet game. Games and Economic Behavior 30, 163–182.
Fountain, J., 2002. Eliciting beliefs from risk averse forecasters using a log scoring rule. Working Paper. University of Canterbury.
Geanakoplos, J.D., Polemarchakis, H.M., 1982. We can't disagree forever. Journal of Economic Theory 28, 192–200.
Glein, I.N., Wallace Jr., J.B., 1974. Probabilistically answered examinations: a field test. The Accounting Review 49, 363–366.
Goeree, J., Holt, C., Palfrey, T., 2003. Risk averse behavior in generalized matching pennies games. Games and Economic Behavior 45, 97–113.
Good, I.J., 1952. Rational decisions. Journal of the Royal Statistical Society B 14, 107–114.
Hambleton, R.K., Roberts, D.M., Traub, R.E., 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.
Huck, S., Weizsäcker, G., 2002. Do players correctly estimate what others do? Evidence of conservatism in beliefs. Journal of Economic Behavior and Organization 47, 71–85.
Kadane, J., Winkler, R., 1988. Separating probability elicitation from utilities. Journal of the American Statistical Association 83, 357–363.
Ledyard J., Hanson, R., Ishikida, T., 2009. An experimental test of combinatorial information markets. Journal of Economic Behavior and Organization 69, 182–189.
McKelvey, R.D., Page, T., 1986. Common knowledge, consensus, and aggregate information. Econometrica 54, 109–127.
McKelvey, R.D., Page, T., 1990. Public and private information: an experimental study of information pooling. Econometrica 58, 1321–1339.
McKelvey, R.D., Palfrey, T., Weber, R., 2000. The effects of payoff magnitude and heterogeneity on behavior in 2×2 games with unique mixed strategy equilibria. Journal of Economic Behavior and Organization 42, 523–548.
Nelson, R.G., Bessler, D.A., 1989. Subjective probabilities and scoring rules: experimental evidence. American Journal of Agricultural Economics 71, 363–369.
Nielsen, L.T., Brandenburger, A., Geanakoplos, J.D., McKelvey, R.D., Page, T., 1990. Common knowledge of an aggregate of expectations. Econometrica 58, 1235–1239.
Nyarko, Y., Schotter, A., 2002. An experimental study of belief learning using elicited beliefs. Econometrica 70, 971–1005.
Ochs, J., 1995. Games with unique mixed strategy equilibria: an experimental study. Games and Economic Behavior 10, 202–217.
Offerman, T., Sonnemans, J., 1998. Learning by experience and learning by imitating successful others. Journal of Economic Behavior and Organization 34, 559–575.
Offerman, T., Sonnemans, J., Schram, A., 1996. Value orientations, expectations, and voluntary contributions in public goods. Economic Journal 106, 817–845.
Offerman, T., Sonnemans, J., van de Kuilen, G., Wakker, P., in press. A truth-serum for non-Bayesians: correcting proper scoring rules for risk attitudes. Review of Economic Studies.
Palfrey, T., Rosenthal, H., 1991. Testing game-theoretic models of free riding: new evidence on probability bias and learning. In: Palfrey, T. (Ed.), Laboratory Research in Political Economy. University of Michigan Press, Ann Arbor, pp. 239–267.
Rutström, E., Wilcox, N., 2008. Stated beliefs versus inferred beliefs: a methodological inquiry and experimental test. Working Paper. Accessed at http://www.class.uh.edu/econ/faculty/nwilcox/papers/stated_vs_inferred_RW.pdf.
Savage, L.J., 1971. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association 66, 783–801.

Seidenfeld, T., 1985. Calibration, coherence, and scoring rules. Philosophy of Science 52, 274–294.
Staël von Holstein, C.-A.S., 1971. An experiment in probabilistic weather forecasting. Journal of Applied Meteorology 10, 635–645.
Staël von Holstein, C.-A.S., 1972. Probabilistic forecasting: an experiment related to the stock market. Organizational Behavior and Human Performance 8, 139–158.
Winkler, R.L., 1968. The consensus of subjective probability distributions. Management Science 15, B61–B75.
Winkler, R.L., 1971. Probabilistic prediction: some experimental results. Journal of the American Statistical Association 66, 675–685.