Economics Letters 111 (2011) 30–33

Contents lists available at ScienceDirect

Economics Letters

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

Incentive effects: The case of belief elicitation from individuals in groups[☆]

Stephanie W. Wang [*]

California Institute of Technology, United States

ARTICLE INFO

Article history:
Received 7 September 2009
Received in revised form 18 November 2010
Accepted 24 November 2010
Available online 4 December 2010

JEL classification:
C91
C72
D83

Keywords:
Belief elicitation
Incentives
Belief aggregation
Experiments

ABSTRACT

Non-incentivized belief elicitation has a negative effect on the belief accuracy of experienced observers predicting choices in 2×2 matrix games. This negative impact extends to the accuracy of group beliefs and revised beliefs after forecasters know each other's initial beliefs.

1. Introduction

A slew of recent studies have included a belief elicitation component because reliable direct measurements of beliefs rather than inferences from choice data that require additional assumptions allow for sharper tests of theories about decision-making and strategic interactions. Some of these studies incentivize the belief reports by paying based on the accuracy of the beliefs via a scoring rule while others do not. However, there have been no studies that cleanly examine how the existence of such incentives affect the accuracy of elicited beliefs.

Our experiment elicits probabilistic beliefs from experienced observers about strategic choices in a 2×2 asymmetric matching pennies game (Nyarko and Schotter, 2002; Palfrey and Wang, 2009) in two treatments: a quadratic scoring rule and a constant payoff not dependent upon belief accuracy. We also elicit beliefs a second time in each round where subjects can revise their forecasts after observing the beliefs of all other members of their group. We report several results that suggest the lack of incentives for accurate beliefs has a substantial negative impact on the accuracy of initial as well as updated individual and group mean beliefs.

1.1. Related literature

Very few studies have looked closely at the impact of the incentive structure for subjects stating beliefs on the beliefs elicited. Gächter and Renner (2006) revisit Croson's (2000) public goods experiment with belief elicitation and add treatments so that they can compare incentivized (beliefs paid based on accuracy) vs. non-incentivized (no payment for beliefs) elicitation. They find that the accuracy of beliefs was significantly higher in the incentivized treatment. However, this comparison was made in an environment in which subjects played the public goods gave as well as stated beliefs thus complicating the interpretation of the incentive effect. Palfrey and Wang (2009) compare beliefs elicited from observers under three different scoring rules, the quadratic, the logarithmic, and the linear, about game play in a previous experiment and find some surprising differences in stated beliefs across treatments even when none were predicted. However, they did not conduct an non-incentivized treatment with no scoring rules as we do here. Whether paying for accuracy of beliefs has a positive effect on accuracy of stated beliefs is unclear given the literature on the possible crowding out effects of extrinsic incentives on intrinsic incentives (Benabou and Tirole, 2003). Furthermore, the quadratic scoring rule we use is no longer proper if the forecaster is risk-averse (hedging towards 50/50) or risk-loving (moving towards extremes) (Offerman et al., 2009). Therefore the scoring rule may distort the stated beliefs in a way that the constant payoff would not.

[☆] I gratefully acknowledge the financial support of The Princeton Laboratory for Experimental Social Science. I thank Roland Benabou, Stephen Morris, Thomas Palfrey, and an anonymous referee for helpful comments.

[*] Division of the Humanities and Social Sciences, Caltech, Pasadena, California 91125, United States.
E-mail address: sweiwang@hss.caltech.edu.

0165-1765/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2010.11.045

2. Experimental design

2.1. Asymmetric matching pennies

Table 1 shows the simple 2×2 asymmetric matching pennies game that was used in the Nyarko-Schotter experiment and in ours as well.

This is a constant sum game with an unique mixed strategy Nash equilibrium. In that equilibrium, both players choose Green with 40% probability and Red with 60% probability.

2.2. Incentive structures

1. *Quadratic Scoring Rule:* The quadratic scoring rule (Brier, 1950) deducts for inaccuracy by subtracting the sum of the square deviations from a constant. Here, the observers forecast a binary outcome: the player being observed either chooses the action Green (G) or the action Red (R). The forecast probabilities placed on the two actions are p_G and p_R, respectively, where $p_G+p_R=1$. Just as in Nyarko and Schotter (2002), we pay our subjects in the quadratic scoring rule treatment a dollar amount proportional to their score:

$$S_G = 1-p_R^2 \text{ if } G \text{ is chosen}$$
$$S_R = 1-p_G^2 \text{ if } R \text{ is chosen}$$

   The quadratic rule is proper under the risk neutrality assumption: a risk-neutral forecaster with true beliefs π maximizes expected score (expected payoff) by reporting $p=\pi$.
2. *Constant Payoff:* In this treatment, we pay the subjects a constant amount per round commiserate with the average payoff in the quadratic scoring rule treatment (0.6). Under this payoff structure, the forecaster is not monetarily rewarded for more accurate beliefs and is given no incentive to reveal true beliefs. Theory is silent on the forecasting behavior in this case since the forecaster would receive the same payoff regardless of the beliefs stated. Forecasters could do anyting from stating the same belief in every round or stating beliefs completely randomly.

We conducted 4 sessions with a total of 32 subjects who were all registered students at Princeton University. Sessions were conducted at Princeton Laboratory for Experimental Social Sciences and all interaction took place through the computers. 8 subjects participated in each session and no subject participated in more than one session. The primary treatment variable was the incentive structure: quadratic scoring rule or constant payoff, with 16 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 2×2 game in Table 1. They played the game repeatedly for five rounds paired with the same opponent. After round 5, they are assigned to the opposite role, are randomly paired 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. The purpose of part 1 of the session was to familiarize the subjects with the strategic task.

In part 2, subjects made "observer" forecasts about the sequence of choices that the row or the column player from seven different matches of the Nyarko-Schotter (NS) experiment made. Four of the subjects (row forecasters) were assigned the task of forecasting behavior of row players and four were assigned the task of forecasting the choices of column players (column forecasters). These roles stayed the same in all of 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. In the first stage of the first round of a match, row forecasters are then asked to report their probabilistic beliefs about the row player in that pair choosing red or green in round six of their NS match and the column forecasters are asked to do the same for the column player simultaneously. They stated beliefs by typing in two positive integers, one for green and one for red, where the two numbers must sum to exactly 100. All row and column forecasters enter their beliefs independently and this concludes stage one. At the beginning of stage two of the same round, all row forecasters are told the stated beliefs of all the other row forecasters, and all column forecasters are told the stated beliefs of all the other column forecasters. The subjects can then change their stated beliefs or keep the same one in stage two. They are asked to re-enter two positive integers, one for green and one for red, that sum up to 100 as their forecast for the same round of play. The revised beliefs of all the row predictors are shown to the row forecasters, and likewise for the column forecasters at the end of stage two.

This concludes one round of one match. The actual choices by the row and column players in round 6 of that NS pair are then reported 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 calculated according to the incentive scheme used for that particular session appears on their screen. (We randomly pick one of the two stages for payment in each round to eliminate incentive distortions) 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 with the same procedures that they used for round 6. 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 to make iterative forecasts for the actions in rounds 8, 9, and 10 of that NS pair. This procedure was then repeated during the session for six other NS pairs. Overall, subjects reported and revised forecasts about a total of 35 rounds of play of the game by 7 different pairs. They were paid the sum of their earnings over all rounds.

Table 1
Game payoffs.

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

3. Results

Before we compare the accuracy of forecasts across the two elicitation conditions, we first look at the difference in their cumulative distribution of forecast extremeness (distance from the 50/50 forecast). We then looked at the correlation of elicited beliefs with actual choice as well as their level of calibration to compare the accuracy of first stage beliefs across the quadratic scoring rule and constant payoff treatments. Finally, we compare the forecasts made in the first stage vs. second stage under our iterative elicitation method where all forecasts within a group was announced to all group members before the second stage forecasts are made. We examine whether the improvements in accuracy of stated beliefs, both at the individual level and the group mean, from the first to second stage differ across treatments.

3.1. Distribution of forecasts

We first compare the distribution of forecasts in the quadratic scoring rule and constant payoff treatments. Specifically, Fig. 1 contains the *extremeness,* the absolute difference from 50, CDF of stated beliefs in our two treatments. We use the 50/50 forecast as the *uninformed* baseline because always stating 50/50 is optimal for a forecaster with uniform prior on [0,1] for the action choice probability. The mean extremeness in the constant payoff treatment, 18.43, is significantly lower than 25

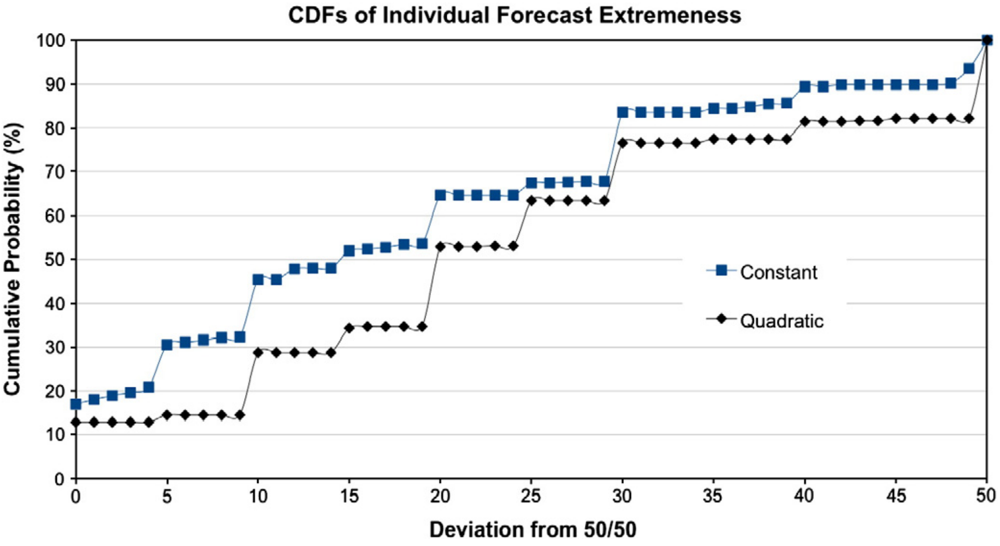

Fig. 1. Individual belief extremeness by treatment.

($p<0.01$)[1] and the CDF is not an uniform distribution. Thus we are able to reject the naive hypothesis that subjects are completely random in stating their beliefs under constant payoff. However, the belief extremeness under the quadratic scoring rule does stochastically dominate that of the constant payoff treatment and the mean, 23.76, is significantly higher ($p<0.1$). Furthermore, the proportion of uninformed beliefs (50/50) is higher under the constant payoff as shown in Fig. 1.

3.2. Initial belief accuracy

To compare belief accuracy across treatments, we first look at the correlations between the stated beliefs about the action choices and the actual choices made by the NS players. This is reported in the first row of Table 2. The stated beliefs under the quadratic scoring rule has significant positive correlation with the actual action choices while the correlation between stated beliefs and action choices in the constant treatment is negative and not significantly different from 0.

To explore the differences further, we use another common performance measure, calibration (Seidenfeld, 1985). We ran separate random effects (Hausman test (1978): $p>0.10$ for both treatments) regressions of the action taken (100 for Green, 0 for Red) on the initial stated probabilistic belief of Green being played for each of our two treatments. The coefficient on the stated belief 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 are completely uninformed. We report these regression coefficients and constants in the second and third rows of Table 2. The coefficient is significantly greater than 0 and the intercept is significantly less than 50 for the quadratic scoring rule treatment. In contrast, the coefficient is negative and not significantly different from 0 and the intercept is not significantly different from 50 for beliefs elicited under a constant payoff.

Both the correlation and calibration measures suggest that stated beliefs are more accurate when the payoffs vary round by round depending on the accuracy of the stated belief as judged by a scoring rule. The beliefs stated under a constant payoff are not easily distinguishable from uninformed beliefs.

3.3. Belief accuracy in the second stage

There were two stages of elicitation in each round where each forecaster was told the stated beliefs of the other three group members in the first stage before stating a belief in second stage about the same action choice. This iterative elicitation method in our experimental design allows us to examine the impact of the iterative elicitation on the accuracy of the second stage stated beliefs compared to those in the initial stage across treatments using the performance measures of the previous section. We do this for both individual beliefs and group mean beliefs.

3.3.1. Updated belief accuracy: Individual

The first row in Table 3 reports the correlation between updated beliefs in the second stage of each round and the corresponding action choice in the NS experiment. The correlation for state beliefs in the constant payoff treatment is more negative in the second stage and now significantly so. The quadratic scoring rule beliefs, on the other hand, are more positively correlated with actual choices after an iteration. These changes is also reflected in the random effects (Hausman test (1978): $p>0.10$ for both treatments) calibration regression results reported in the second and third rows of Table 3. The calibration coefficient is now significantly negative for updated beliefs under constant payoff (recall that the coefficient should be 1 under perfect calibration). In contrast, the calibration coefficient of updated beliefs is a small improvement over that of the initial beliefs under the quadratic scoring rule.

3.3.2. Updated belief accuracy: Group mean

Next we turn to comparing the accuracy of the group mean belief in the first stage versus the second stage. Table 4 presents the group mean belief correlations with actual choice as well as the random effects (at the group level; Hausman test: $p>0.10$) calibration regression results just as in Tables 2 and 3. The first two columns contain the first and second stages of the constant payoff treatment while the third and forth columns are the first and second stages under the quadratic scoring rule.

We first note, perhaps not surprisingly, that the constant payoff and quadratic rule group mean beliefs have worse and better accuracy respectively compared to the individual beliefs. The calibration coefficient

Table 2
Individual belief accuracy in first stage by treatment.

	Constant	Quadratic
Correlation	−0.049	0.077
Calibration regression coefficient	−0.10 (0.087)	0.15* (0.069)
Calibration regression constant	54.31(4.81)	42.18$^+$(4.59)

*: Significantly greater than 0 ($p<0.05$).
$^+$: Significantly less than 50 ($p<0.05$).

[1] Unless otherwise noted, all statistical tests are done with standard errors clustered at the subject level.

Table 3
Individual belief accuracy in second stage by treatment.

	Constant	Quadratic
Correlation	−0.075	0.099
Calibration regression coefficient	$-0.16^{\dagger}$ (0.090)	0.19^{**} (0.075)
Calibration regression constant	57.16(4.92)	40.39^{+}(4.60)

Significantly greater than 0 (*: $p<0.05$; **: $p<0.01$).
†: Significantly less than 0 ($p<0.05$).
+: Significantly less than 50 ($p<0.05$).

Table 4
Individual belief accuracy in second stage by treatment.

	Constant (1^{st})	Constant (2^{nd})	Quadratic (1^{st})	Quadratic (2^{nd})
Correlation	−0.096	−0.12	0.14	0.16
Calibration coefficient	−0.39 (0.34)	−0.44 (0.30)	0.41 (0.27)	0.43^{*} (0.24)
Calibration constant	68.62(17.52)	70.77(15.30)	29.84^{+}(13.62)	29.57^{+}(12.18)

*: Significantly greater than 0 ($p<0.05$).
+: Significantly less than 50 ($p<0.05$).

is even more negative and the constant even higher for the constant payoff treatment when the group mean belief is used. A crowd of more accurate forecaster is “wise" while the crowd of less accurate forecasters is “confused." In addition, while these accuracy measures do not change much for the second stage updated beliefs in this treatment, they do move toward even greater inaccuracy. The group beliefs under the quadratic scoring rule have a higher calibration coefficient than the individual beliefs and a constant that is much closer to 0 (the constant value under perfect calibration). Furthermore, the coefficient becomes significantly positive for the second stage beliefs even though the magnitude again shifts very little.

We find that the differences in individual belief accuracy between the treatments are magnified when group mean belief accuracy is measured. Furthermore, there is suggestive evidence that the second stage beliefs, be they individual or group, are not improving if not becoming worse in terms of accuracy for the constant payoff treatment. One possible explanation is that forecasters have a harder time of interpreting and incorporating the beliefs of others in their updating when those beliefs may very well be distorted by incentives or lack thereof (Lichtendahl and Winkler, 2007).

4. Conclusion

We report several findings. The distribution of deviation from the 50/50 uninformed forecast under the constant payoff is stochastically dominated by that under the quadratic scoring rule. The initial individual stated beliefs in the constant payoff treatment are not positively correlated with actual action choices while those under the quadratic scoring rule are. Furthermore, the individual beliefs are worse calibrated in the constant payoff treatment. We also find that this difference in accuracy across treatments is even more pronounced when we look at the initial group mean belief. Finally, neither the individual or belief mean belief accuracy changes very much after subjects have the chance to update their beliefs. To the extent that they do change, the constant payoff condition does worse while there is slight improvement in the quadratic scoring rule condition.

Our findings show that stated beliefs are less accurate and closer to uninformed when belief accuracy is not being rewarded. For a moderately difficult prediction task such as ours, forecasters likely need to be incentivized to expend the necessary effort to form good predictions. Our results caution against not using proper incentives in belief elicitation tasks out of concerns for introducing complexity or distortions. We also find suggestive evidence for a negative externality of non-incentivized belief elicitation, namely the inability of other forecasters to make inferences and correctly Bayesian update when faced with noisy beliefs elicited under flat monetary incentives. Given the importance of eliciting precise, deliberated beliefs from forecasters for decision-making and theory-testing, providing them with enough incentives to state such beliefs should be a priority.

References

Benabou, R., Tirole, J., 2003. Intrinsic and extrinsic motivation. Review of Economic Studies 70, 489–520.

Brier, G., 1950. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review 78, 1–3.

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.

Gächter, S., Renner, E., 2006. The effects of (incentivized) belief elicitation in public good experiments, Working Paper.

Hausman, J.A., 1978. Specification tests in econometrics. Econometrica 46, 1251–1271.

Lichtendahl, K.C., Winkler, R.L., 2007. Probability elicitation, scoring rules, and competition among forecasters. Management Science 43, 1745–1755.

Nyarko, Y., Schotter, A., 2002. An experimental study of belief learning using elicited beliefs. Econometrica 70, 971–1005.

Offerman, T., Sonnemans, J., van de Kuilen, G., Wakker, P.P., 2009. A truth-serum for non-Bayesians: correcting proper scoring rules for risk attitudes. Review of Economic Studies 76, 1461–1489.

Palfrey, T.R., Wang, S.W., 2009. On eliciting beliefs in strategic games. Journal of Economic Behavior and Organization 71, 98–109.

Seidenfeld, T., 1985. Calibration, coherence, and scoring rules. Philosophy of Science 52, 274–294.