The Binarized Scoring Rule*

Tanjim Hossain
University of Toronto

Ryo Okui
Kyoto University

July 2012

Abstract

We introduce a simple method for constructing a scoring rule to elicit an agent's belief about a random variable that is incentive compatible irrespective of her risk-preference. The agent receives a fixed prize when her prediction error, defined by a loss function specified in the incentive scheme, is smaller than an independently generated random number and earns a smaller prize (or a penalty) otherwise. Adjusting the loss function according to the belief elicitation objective, the scoring rule can be used in a rich assortment of situations. Moreover, the scoring rule can be incentive compatible even when the agent is not an expected utility maximizer. Results from our probability elicitation experiments show that subjects' predictions are closer to the true probability under this scoring rule compared to the quadratic scoring rule.

Keywords: Scoring rule, belief elicitation, randomization, prediction, risk-preference, laboratory experiment.

JEL classification: C91, D81, D83, D84.

*We thank Robin Chark, Eric Set, and Keigo Inukai for their excellent research assistance, the Managing Editor (Imran Rasul), four anonymous referees, Phil Dybvig, Haluk Ergin, Glenn Harrison, Atsushi Kajii, Sachar Kariv, Nicolas Lambert, David Levine, Colin Stewart, Norio Takeoka, and seminar participants at University of Toronto, Washington University St. Louis, Yokohama National University, Nagoya University, Hitotsubashi University, the spring meeting of the Japanese Economic Association, and the winter meetings of the Econometric Society for helpful comments. We are grateful to Sudipto Dasgupta who initially got us interested in the problem and Yoichi Hizen who kindly helped us in using the Group Experiment Lab in the Center for Experimental Research in Social Science of Hokkaido University. We gratefully acknowledge the financial support from Kyoto University, SSHRC through grant no. 489160, and HKUST through grant no. RPC06/07.BM04. Please direct all correspondence to Tanjim Hossain at tanjim.hossain@utoronto.ca.

1 Introduction

We introduce a parsimonious but general way of constructing a scoring rule to elicit an agent's beliefs about a random variable that is incentive compatible irrespective of her utility function. The main characteristic of this incentive scheme is that it creates a binary function that determines whether the agent will receive a fixed reward based on her performance using an independently drawn random number. Specifically, after observing information about a random variable, the agent reports her prediction of some feature of that variable incorporating her beliefs. She receives the reward if her prediction error, defined by a loss function, is smaller than an independently drawn random number from a uniform distribution and earns a smaller prize or a penalty otherwise. In other words, she receives the reward with a probability which is determined by the realized value of the loss function. Her objective, thus, becomes maximizing the probability of winning the reward. This probability is proportional to the expected loss. Under relatively weak assumptions, she can maximize her expected (or non-expected) utility by taking the action that minimizes the expected value of the loss function. We term this incentive scheme the *Binarized Scoring Rule* or *BSR*.

Eliciting an economic agent's beliefs concerning a probabilistic event (which may or may not be objectively specified) is an important problem. Many have suggested scoring rules to incentivize truthful communication of agent's beliefs. Many of these mechanisms, such as the proper scoring rules suggested by Savage (1971) or the methods of promissory notes by De Finetti (1974), however, lead to agents reporting their true belief only under risk-neutrality. For example, we may want to know an agent's belief about the probability of the occurrence of a particular event. A well-known method is the quadratic scoring rule (QSR) of Brier (1950). Consider a random scalar variable X about which the agent holds some belief. Suppose we want to elicit the probability of the event that x, which is the realized value of X, exceeds some number n using the QSR. First, the agent reports p as the probability of the event occurring. Then, after X is realized, she is paid $1-(1-p)^2$ if $x > n$ and $1-p^2$ otherwise. It is easy to see that reporting her true belief maximizes her expected utility if she is risk-neutral. However, if she is risk-averse, reporting a number between 0.5 and her true belief may yield a higher expected utility. Generally speaking, under risk-aversion, the marginal utility of the monetary payment to the agent confounds the effect of her beliefs making belief elicitation difficult.

Our method modifies a proper scoring rule, which is valid under risk-neutrality, by binarizing it in a way that the agent gets a fixed reward if the realized score, calculated

using the relevant loss function, is lower than a number drawn from a uniform distribution. The realized score determines the probability of the reward, but not the size of the reward. For example, to elicit the probability of the realized value of X being greater than n, we binarize the QSR by using the loss function that equals $(1-p)^2$ if $x > n$ and p^2 if $x \leq n$. That is, the agent receives a fixed reward when the realization of a random number independently drawn from the uniform distribution on $[0,1]$ is above $(1-p)^2$ if $x > n$ and above p^2 if $x \leq n$. In other words, the agent receives the reward with probability $1-(1-p)^2$ if $x > n$ and with probability $1-p^2$ otherwise. Since she prefers receiving the reward over not receiving it, she will maximize the probability of receiving the reward under this scoring rule. Interestingly, the agent's optimization problem, independent of her risk-preference, is the same as that under the QSR for risk-neutral agents. Thus, this scheme is incentive compatible no matter whether the agent is risk-neutral, risk-averse, or risk-seeking. By turning the incentive scheme into a lottery where the sizes of the rewards do not depend on p, this scoring rule induces risk-neutral behavior.

We can adjust the loss function according to our goal for eliciting the agent's beliefs. We do not just provide a scoring rule, rather we suggest a method to devise appropriate scoring rules according to the belief elicitation objective. Instead of the probability of the realized value of X being greater than n, suppose we want to elicit the probability distribution of X (let us assume that X can take one of N possible values). Then, we can employ the loss function $\sum_{i=1}^{N}(\mathbf{1}_i - p_i)^2$ where $\mathbf{1}_i$ is an indicator function that equals 1 if X takes the i-th value and 0 otherwise and p_i is the reported probability of X taking the i-th value. If we are interested in eliciting the expected mean of X, we can use the loss function of $(x-m)^2$ where m is the reported expected value of X. To elicit the expected median, we can use the loss function $|x - md|$ where md is the reported expected median of X. We can elicit the α% quantile by choosing the loss function to be $|x-q|\left(\alpha \mathbf{1}_{\{x>q\}} + (100-\alpha)\mathbf{1}_{\{x\leq q\}}\right)$ where q is the reported α% quantile. Gneiting and Raftery (2007) provide many other examples of scoring rules that can be used as loss functions.

The binarized scoring rule may even be used when the agent's decision mechanism is described not by the expected utility theory. Recall that as the agent gets a higher utility from the larger reward, an expected utility maximizing agent chooses her action to maximize the probability of winning the larger reward. Now suppose her preferences do not satisfy expected utility theory but, still, is such that she prefers a binary lottery that puts a higher probability on the larger reward to one with a lower probability. In that case,

she still prefers maximizing the probability of winning the larger reward. As a result, the binarized scoring rule is incentive compatible as long as the agent's preferences satisfy monotonicity with respect to first order stochastic dominance restricted to the edges of the simplex.[1]

To illustrate how the binarized scoring rule can be used in practice, we also present results from two sets of experiments. As a benchmark, we compare its performance with that of the quadratic scoring rule. One experiment is designed to elicit the probability of the occurrence of a specific event and the other to elicit the expected value of a random variable. In both experiments, the distribution of the relevant random variable is specified so that beliefs can be specified objectively. We investigate how closely subjects report the "correct" probability or mean under the BSR in comparison to the QSR. In the experiment where we elicit the probability of an event, the BSR performs better than the QSR in this experiment. The superior performance of the BSR can be solely attributed to risk-averse subjects. On the other hand, the BSR and the QSR perform equally well in the experiment in which we elicit the mean. These results are consistent with theoretical predictions.

The rest of the paper is structured as follows. The following subsection relates our paper to the existing theoretical and experimental literature on belief elicitation. Section 2 presents the binarized scoring rule and the theorems that characterize this scoring rule. Section 3 describes the laboratory experiments and discusses the results from these experiment. Section 4 concludes. All proofs are in the appendix.

1.1 Relation to the literature

Theoretically, our method formalizes and generalizes the intuition in Smith (1961) and Roth and Malouf (1979) who basically suggest paying agents with "probability currency" for a binary lottery instead of direct monetary payments to induce risk-neutral behavior.[2] We show that this idea can be utilized in a host of belief elicitation problems as long as there exists a proper scoring rule that is incentive compatible under risk-neutrality.[3] We

[1] We thank an anonymous referee for pointing this out.

[2] Berg, Daley, Dickhaut and O'Brein (1986) explore this idea more formally. This intuition has also been used in many other experimental settings ranging from auctions to battle of sexes games. See, Selten, Sadrieh, and Abbink (1999) for a survey. An interesting application is by Laury, McInnes, and Swarthout (2012) who develop a method to elicit discount rates without assuming the form of the utility function.

[3] A limitation of our mechanism is that it does not work when the agent has a personal stake on the event. That is, she receives a reward that is different from the monetary reward given by the elicitor, that depends on the realization of the random variable and is not observable to us (see Kadane and Winkler (1988)). In fact, Karni and Safra (1995) demonstrate that, without knowing the utility function

also extend the results to a setting under the non-expected utility paradigm.

For specific problems, several previous studies developed scoring rules that do not require assumptions on utility function. Bhattacharya and Pfleiderer (1985) show that the quadratic scoring rule can elicit the mean of a random variable under the belief if the distribution is symmetric and the agent is a (weakly) risk-averse expected utility maximizer. For eliciting the probability of a certain event, Allen (1987) proposed a randomized scoring rule where the agent receives a fixed reward if the reported probability is below a randomly drawn number. However, this random number is drawn after the event is realized and the distribution from which the number is drawn depends on whether the event occurred or not. When there are only two possible outcomes, this method is equivalent to the BSR based on a quadratic loss function. However, when there are three or more outcomes, this method involves an additional layer of randomization.

Recently, Karni (2009) proposed a more complicated mechanism that involves two layers of randomization.[4] First the agent reports her prediction of the probability of the event occurring. If this probability is below a randomly drawn number then the agent receives a binary lottery that gives the preferred reward with the probability equaling the random number. Otherwise, she receives a binary lottery that gives the preferred reward if the event in question occurs. There are several differences between the BSR and the Karni mechanism. First, under the BSR, we offer the subject a lottery whose winning probability is determined through the experimental procedure but does not directly depend on the probability of the event we are interested in. Second, the BSR method has only one layer of randomization. We think that this feature is important in applications because it makes the procedure simple. Lastly, the BSR can be applied in many different situations in which we are interested in properties of random variables different from probabilities. On the other hand, it is not clear how to extend the Karni mechanism to elicit other features than probabilities such as the expected value of a variable. Qu (2012) cleverly extends the Karni mechanism to elicit the entire probability distribution of random variables that may not be binary. Note that Qu (2012) introduces another layer of randomization and his method is further complicated than the Karni mechanism.

Independently of our work, Schlag and van der Weele (2009) developed a probabilis-

completely, there is no scoring rule that elicits the probability of an event if the agent has a stake on the event. Thus, this problem is not only for our scoring rule, but also for any scoring rule. This may not be critical in a laboratory experiment because it is unlikely that a subject has a personal stake related to a random variable generated in a laboratory.

[4]Grether (1981) and Holt (2007, Chapter 30 Appendix and Question 6) independently suggested the same mechanism in more heuristic manners.

tic version of the quadratic scoring rule for eliciting probability using a method which is essentially the same as ours but restricted to the case where the support of the loss function is *finite*. The BSR, on the other hand, can be used even when the loss function is unbounded and under more general settings. We also ran laboratory experiments to measure the performances of the BSR. Alternative approaches to elicit beliefs under risk-aversion also include those by Andersen, Fountain, Harrison, and Rutström (2010) who jointly estimate agent's risk attitude and subjective probability and Offerman, Sonnemans, van de Kuilen, and Wakker (2009) who provide a way to correct the reported probability to recover subjective probability.[5] These methods are usually devised for a specific belief elicitation problem and it is not straight-forward to extend them to other problems.

Selten, Sadrieh, and Abbink (1999) contend that rewards determined by a lottery do not induce risk-neutral behavior in subjects in their laboratory experiment and that those schemes perform much worse than do money prizes.[6] Our experimental results alleviates such concerns by finding that the BSR outperforms the QSR. There has been a number of experimental investigations to elicit belief of agents in the laboratory—some using deterministic and other using probabilistic scoring rules.[7] There are fewer studies that experimentally compare the performances of various scoring rules. Among them, most closely related to our paper are Andersen, Fountain, Harrison, and Rutström (2010) who compare the QSR and a linear scoring rule, Hao and Houser (2012) who compare two variations of the mechanism by Karni (2009), Hollard, Massoni, and Vergnaud (2010) who found methods similar to the one proposed by Karni (2009) and simply asking subjects about beliefs outperform the QSR, and Trautmann and van de Kuilen (2011) who run a horse race of different elicitation methods and find that their performances do not vary much.[8] Evaluating performance of the BSR relative to that of the QSR is, thus, another contribution of our paper to the literature.

[5] Andersen, Fountain, Harrison, Hole, and Rutström (2012) extend the joint estimation approach to consider cases in which subjective probabilities themselves are random.

[6] Contrary to their findings Harrison, Martínez-Correa, and Swarthout (2012) find that binary lotteries move actions of risk-averse subjects closer to risk-neutrality.

[7] See McKelvey and Page (1990), Möbius, Niederle, Niehaus, and Rosenblat (2007), Holt and Smith (2009) for studies that use probabilistic scoring rules. The introduction of Offerman, Sonnemans, van de Kuilen, and Wakker (2009) provides a nice survey of papers that use scoring rules in various different fields.

[8] A different stream of research asks whether a scoring rule can correctly recover the induced belief. See, for example, Hurley and Shogren (2005) and Blanco, Eugelmann, Koch, and Normann (2010). Alternatively, Hurley, Peterson, and Shogren (2007) study whether scoring rules work better than prediction based elicitation of Grether (1980, 1992).

2 The Binarized Scoring Rule

Now we formally present the binarized scoring rule in quite a general setting. Let X be a random variable which can be a scalar, a vector, or even an infinite-dimensional stochastic process. Suppose we want to elicit the agent's belief about some characteristic of X. Accordingly, the agent reports $\theta \in \Theta$ as the predicted value of this characteristic. Here θ can be a scalar, a vector, or a function. Let the loss function $l(X, \theta)$ be a scalar valued function of the realized value of X and θ. We denote the expectation operator with respect to the distribution of some variable Z by E_Z and make the following assumption on l.[9]

Assumption 1. *i) The realized value of the loss function is non-negative,* $l(X, \theta) \geq 0$ *for all* X, θ.
ii) The expression $\arg\min_{\theta \in \Theta} E_X [l(X, \theta)]$ *is well-defined.*

Our goal is to devise an incentive structure or scoring rule under which the agent will report a value of θ that minimizes $E_X [l(X, \theta)]$, irrespective of the exact form of her preference. The time-line of our proposed mechanism is as follows:

1. The agent reports θ to the principal.
2. X is realized.
3. The principal draws K from $U[0, \overline{K}]$ (uniform distribution whose support is $[0, \overline{K}]$ for a positive number $\overline{K}$) independently of the realization of X or the reported θ.
4. The agent receives reward A if $l(X, \theta) < K$ and reward B otherwise where she prefers A to B.

We call this mechanism the binarized scoring rule (BSR) because it creates a binary lottery whose outcome depends on the realizations of the loss function and the random variable K. Here, the agent receives preferred reward A with the probability that the loss is less than K. Let $P(\theta)$ denote this probability so that $P(\theta) = E_K E_X[\mathbf{1}_{\{l(X,\theta)<K\}}]$. Here $\mathbf{1}_{\{l(X,\theta)<K\}}$ is the indicator function that equals 1 if $l(X, \theta) < K$ and 0 otherwise. The

[9] We can describe the setting with a more general and mathematically rigorous representation. Let Ω be some set and $\mathcal{F}$ be a σ-field of subsets of Ω and $\mathbb{P}$ be a probability measure such that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Let the loss function $l : \Omega \times \Theta \to \mathbb{R}$ be a measurable function for any θ where Θ is a parameter space. The principal then tries to let an agent report a value of θ that minimizes the expected loss: $\int_\Omega l(\omega, \theta) d\mathbb{P}(\omega)$. Both the principal and the agent observe $l(\omega, \tilde{\theta})$, where $\tilde{\theta}$ is the reported value of θ, after the agent reports $\tilde{\theta}$. The results of this paper hold under this general setting.

rewards A and B are chosen such that the agent prefers A to B. When we use monetary rewards, setting $A > B$ would suffice as long as the agent likes money, which, we believe, is hardly a controversial assumption. On the other hand, some non-monetary rewards can be used too as long as we know that A is preferred to B.

To understand the motivation of the mechanism and how it works, let us consider the case in which the agent is an expected utility maximizer with utility function u. When the agent is risk neutral so that $u(x) = x$, then paying her the amount $-l(X, \theta)$ after X realized provides sufficient incentives for her to report the value of θ that minimizes $E_X[l(X, \theta)]$. However, for other risk preferences, the value of θ that maximizes the expected utility $E_X[u(-l(X, \theta))]$ is usually different from the value of θ that minimizes $E_X[l(X, \theta)]$. This problem provides the motivation for developing an incentive scheme that is incentive compatible irrespective of the form of u. The expected utility under the BSR is $E_K E_X[u(A\mathbf{1}_{\{l(X,\theta)<K\}} + B\mathbf{1}_{\{l(X,\theta)\geq K\}})] = E_K E_X[\mathbf{1}_{\{l(X,\theta)<K\}}]u(A) + E_K E_X[\mathbf{1}_{\{l(X,\theta)\geq K\}}]u(B) = P(\theta)u(A) + (1 - P(\theta))u(B)$. When $u(A) > u(B)$, maximizing the expected utility becomes equivalent to maximizing $P(\theta)$. If $l(X, \theta) \leq \overline{K}$ holds for any X and θ, then this probability equals $1 - E_X[l(X, \theta)]/\overline{K}$. Therefore, the agent would report θ that minimizes $E_X[l(X, \theta)]$. The keys here are that the BSR creates a binary lottery and that the utility is increasing in the probability of receiving the reward A.

The BSR reduces the incentive scheme to an environment where two alternatives are awarded with differing probabilities. Thus, the objective becomes maximizing the probability of receiving the more attractive reward A. As a result, the scheme is incentive compatible even when the agent's decision is not represented by expected utility maximization. We only assume that between two binary lotteries with rewards A and B, she prefers the one with the higher probability of winning A. To state this result formally, we start with describing our setting on the preference of the agent. Let $L(p)$ denote a binary lottery that gives A with probability p and B with probability $1 - p$ so that $L(p) = [A, p; B, 1 - p]$. Let V be the real-valued preference functional on the set of lotteries that can be written as $L(p)$ for some p. The essential assumption here is that $V(L(p)) > V(L(p^*))$ if and only if $p > p^*$. This is a weak version of so-called *monotonicity with respect to stochastic dominance* by Machina and Schmeidler (1992, p754).[10] Many non-expected utility theories, such as the theory of Machina (1982) and the rank-dependent expected utility theory of Quiggin (1981), satisfy this property. Note

[10] In Machina and Schmeidler (1992), monotonicity with respect to stochastic dominance is a part of the definition of probabilistically sophisticated non-expected utility maximizer. They also provide an axiomatic foundation of probabilistic sophistication.

that, we basically assume that the principal chooses A and B in a way that this property is satisfied for binary lotteries involving A and B. We do not assume anything for preferences not concerning prospects A and B. We also do not need to make any assumption on the initial wealth level of the agent.

The following theorem shows that the value of θ that maximizes the expected utility under our scoring rule is the same as the minimizer of $E_X[l(X,\theta)]$ which we would like to elicit. We make an additional assumption that the loss function is bounded, which we discuss later.

Theorem 1. *Suppose that Assumption 1 holds. Assume that V is monotone with respect to stochastic dominance in the sense that $V(L(p))) \geq V(L(p^*)))$ for $p \geq p^*$ and the inequality is strict if $p > p^*$. Assume that $l(X,\theta) \leq \overline{K}$ for any θ and X. Then,*

$$\arg\max_{\theta\in\Theta} V(L(P(\theta))) = \arg\min_{\theta\in\Theta} E_X[l(X,\theta)].$$

The decision made by an agent under the BSR is equivalent to choosing a lottery from the set of lotteries which is indexed by θ. Because K is uniform and l is always in the support of K, the probability $P(\theta)$ is negatively affine to the expected value of l. Therefore, maximizing the preference is equivalent to minimizing the expected loss. When there are multiple maximizers of the preference functional, the theorem implies that all of them are minimizers of the expected loss.

The conditions on the preference for the preceding theorem are weak and can be satisfied by many theories of decision under uncertainty. Nonetheless, they implicitly impose several restrictions. We assume that the agent is "probabilistic sophisticated."[11] Thus, it is not clear how the BSR works when probabilistic sophistication is violated. The theorem also relies on the *reduction of compound lotteries axiom*. Harrison, Martínez-Correa and Swarthout (2012) provide an excellent review of the literature and an experimental investigation of this axiom.[12]

The limitation of the theorem that may be important in practice is the assumption that $l(X,\theta)$ is bounded by $\overline{K}$. Nevertheless, under many circumstances, this boundedness assumption can be relaxed. Additional assumptions required to relax the boundedness assumption are specific to the loss function. In some experiments in this paper, we

[11] See Machina and Schmeidler (1992), Grant (1995), and Chew and Sagi (2006) for axiomatic foundations of probabilistic sophistication.

[12] This axiom is possibly violated in the face of dynamic inconsistency. See Machina (1989) for a review. This axiom is also problematic in the axiomatic foundation of social welfare function as originally illustrated by Diamond (1967). It is therefore important to examine the validity of our proposed mechanism using experimental or real data.

consider the case in which we want to minimize the mean squared error concerning the realized value of the scalar random variable X. Hence, the relevant loss function is quadratic, $l(X,\theta) = (X-\theta)^2$. We show that if the distribution of X has a light tail, the value of θ that maximizes the preference approximates the mean when the loss function is quadratic. Moreover, if the distribution is symmetric, then reporting the mean maximizes the preference. These results are summarized in the following theorem.

Theorem 2. *Suppose that Assumption 1 holds with $l(X,\theta) = (X-\theta)^2$ where X is a scalar that is described by the density function $f(a)$ and has finite second moments. Moreover, $|a|^{2+\delta} f(a) \to 0$ for some $\delta > 0$ as $a \to \infty$ and as $a \to -\infty$. As $\overline{K} \to \infty$,*

$$\arg\max_{\theta\in\Theta} V(L(P(\theta))) \to E_X[X],$$

where $P(\theta) = E_K E_X[\mathbf{1}_{\{(X-\theta)^2<K\}}]$. Furthermore, when $f(a)$ is symmetric around the mean, there exists a finite $\widetilde{K}$ such that for any $\overline{K} > \widetilde{K}$,

$$\arg\max_{\theta\in\Theta} V(L(P(\theta))) = E_X[X].$$

Proposition 2 in Bhattacharya and Pfleiderer (1985) states that the QSR can be used to elicit the mean if the agent is a weakly risk averse expected utility maximizer and the distribution of X is symmetric. We find that the BSR is more widely applicable than the QSR as it is incentive compatible even for risk-loving agents or non-expected utility maximizers and it is so at the limit as $\overline{K}$ approaches infinity even if the underlying distribution is not symmetric. An important practical question is how to choose $\overline{K}$. If the underlying distribution is symmetric, $\overline{K}$ only needs to satisfy the second order condition of the minimization problem. This requirement is easily satisfied. For example, if f is the standard normal density function, then $\overline{K} > 0.3$ is sufficient. When the distribution is not symmetric, then the value of $\overline{K}$ determines the distance between what exactly we can elicit and the true mean. For example, when f is the χ^2_1 density function and $\overline{K} = 50$, the minimizer is around 0.962 so the distance from the true mean, 1, is around 0.038.

3 Experimental Illustration of the BSR

We applied our scoring rule in an experimental framework to analyze belief elicitation about the realized value of a random variable. Using the incentive schemes suggested in Theorems 1 and 2, we ran two sets of experiments in which we elicit subjects' beliefs about certain aspects of a random variable. In the first set, which we call the P-experiment, we elicit subjects' predictions of the probability that a ball drawn from an urn is of

some specified color using both the binarized and the quadratic scoring rules. In the second set, called the M-experiment, a subject gets a number of signals about the realized value of a random variable and then reports her estimate of the realized value. We use relatively simple but non-trivial exercises in these two experiments so that the impact of the incentive schemes is not too confounded by the complexity of the exercise.

The P-experiment was run in Hokkaido University, Japan in July 2010 and the M-experiment was run in Hong Kong University of Science and Technology in June 2009. The experiments were programmed and conducted using the software *z-Tree* developed by Fischbacher (2007). All subjects were undergraduate students of the respective institutions and were recruited using databases of students willing to participate in economic experiments.

3.1 P-Experiment

In the P-experiment, 153 subjects participated in 12 sessions. The subjects were asked to report their prediction concerning the event that a ball randomly drawn from an urn with 100 balls of three different colors—red, black and blue—was of a certain color or was *not* of a certain color. To ensure that the subjects were aware of the concept of probability, we first reviewed the probability rules in this simple setting. Then, they participated in 10 unpaid practice periods where they reported their prediction under both scoring rules. We informed them of the outcome of the draw and how much they would have earned for their predictions after every practice period. Next, they participated in 2 paid periods. They were paid using the BSR in one period and using the QSR in the other. Which incentive scheme would be used in the first period and, for each period, the color composition of the urn, and the event which the subject had to predict were randomly decided. The subjects were clearly informed of the color composition and the relevant event at the beginning of each period. Thus, they knew the true objective probability of the event in each period. To alleviate the concern that a subject's income in the first paid period may affect her choices in the second, they were informed of the outcomes of the draw and their earnings from both periods only after the second paid period.

In the paid period under the BSR, a subject's optimal choice is to report the objective probability of the event happening irrespective of her risk-preference. However, her optimal prediction under the QSR depends on her risk-preference. Before participating in the practice and paid periods, subjects also participated in a 5-period round in each period of which they reported their certainty equivalent for a lottery. The lottery

involved receiving JPY 10 as the low prize and JPY 50 or JPY 100 as the high prize. The probability of winning the larger prize varied between 0.20 and 0.90. The certainty equivalent was elicited using the BDM mechanism proposed by Becker, DeGroot, and Marschak (1964).[13] At the end of a session, subjects were paid in cash for the 2 paid periods and a randomly chosen period from the risk-preference elicitation round. The sessions were conducted in Japanese and the experimental rules were described using PowerPoint presentations. Subjects were provided with handouts illustrating the relationship between their prediction and the payment.[14] In total, subjects spent around an hour in the laboratory on average and the average payment to a subject was around JPY 1745 (USD 21).

Each subject started her session with an initial endowment of JPY 1000. Suppose a subject was asked to report her prediction that the color of the drawn ball (from an urn with 100 balls) was red and she entered an integer P between 0 and 100 (inclusive) as her prediction. Thus, the reported probability of the event happening is $P/100$. Let us define the subject's squared error to be $(1 - P/100)^2$ and $(P/100)^2$ if the drawn ball turned out to be red and not red, respectively. Under the BSR incentive scheme in the paid round, JPY 500 was added to the subject's endowment if her squared error was below a random number K generated from a uniform distribution on $[0, 1]$. If the squared error was below K, then JPY 300 would be taken away from her endowment. Under the QSR incentive scheme, the subject received JPY $500 - 800sqe$ where sqe stands for the squared error. If a subject reported the true objective probability as her prediction (the optimal choice under risk-neutrality), her expected earning would be the same under both incentive schemes.

3.1.1 Results

To compare subjects' performances under the two scoring rules, we compute the measure NSD—the negative of the square of the difference between the reported number and the true probability.[15] Under the BSR, reporting the objective or true probability of

[13]The BDM mechanism is not valid if the agent is not an expected utility maximizer (Holt, 1986). Assuming that compound lotteries can be reduced to simple lotteries, Karni and Safra (1987) show that the BDM method is valid if and only if the expected utility hypothesis holds. To our knowledge, there is no mechanism that elicits the certainty equivalence of a lottery without relying on the expect utility assumption. Nonetheless, we believe that the BDM method provides us with some information about an agent's risk preference.

[14]These and other instructions can be supplied by the authors upon request.

[15]We express NSD in percentage terms in order to make the results easier to present in a table format. Suppose the true and the reported probabilities are $\pi/100$ and $P/100$, respectively. Then, the NSD equals $-(\pi - P)^2$.

the specified event happening maximizes a subject's expected utility, independent of her risk preference. Under the QSR, however, it maximizes a subject's utility only if she is risk-neutral. Of course, we do not expect the QSR to induce risk-neutrality. Rather, as risk-averse subjects are likely to make predictions relatively further from the true probability under the QSR, we use it as a benchmark to evaluate BSR's performance in eliciting the true probability.

For our analysis, we exclude the outlier data points. Specifically, we exclude observations that did not satisfy a criterion we call "betweenness," allowing for some randomness in reporting, and those with the lowest 5% NSD. We consider that an observation satisfies betweenness if the reported probability is between the true probability and 0.5 (inclusive). To allow for small deviations from theory in reporting the probability, we only exclude an observation that does not satisfy betweenness if the distance between the true and reported probabilities is greater than 0.2. However, the empirical results stay qualitatively unchanged whether we exclude observations with the distance between the true and reported probabilities greater than 0, 0.05, 0.1, or 0.15, instead. A typical excluded observation is to report 0.74 when the true probability is 0.26. Such observations are likely to have been caused by simple mistakes and it is appropriate to exclude them. There are 30 excluded observations by 28 subjects (observations from both periods were excluded for two subjects) —16 observations under the BSR and 14 under QSR. Thus, we have a panel with 276 observations from 151 subjects. We assume that the observations are independent across subjects but are potentially correlated over periods.

The first column of Table 1 presents descriptive statistics of NSD for the P-experiment. This suggests that the subjects reported probabilities relatively closer to the objective or true probability under the BSR. To confirm this result, we regress NSD on BSR, which is a dummy variable that indicates when the BSR is used, and other variables to examine the effect of incentive scheme on reported predictions. Table 2 summarizes the results of these regressions.[16] Column (1) present regression result with all 276 observations. The prediction error falls more by one-third when we switch from the QSR to the BSR. The result is robust to controlling for personal characteristics such as gender, major, or class level of the subjects. Next we utilize the risk-preference elicitation round to analyze the performances of subjects based on their risk attitude. Columns (2) to (4) present similar regressions as those in column (1), but with only risk-averse, risk-neutral, and risk-loving

[16] We always present heteroskedasticity and autocorrelation robust standard errors, which allows arbitrary heteroskedasticity and correlation within an individual but assumes independence across subjects (Arellano, 1987).

subjects, respectively.[17] We find that the BSR performs better than the QSR for risk-averse subjects. The benefit of the QSR is not statistically significant for risk-neutral or risk-loving subjects. Recall that, the QSR is incentive compatible for risk-neutral agents. For risk-loving subjects, the directionality of the prediction under the QSR is rather complicated and depends critically on the specific utility function. Nevertheless, subjects will have a tendency to report relatively extreme probabilities.

As risk-averse agents should choose predictions closer to 0.5 under the QSR, the difference between the two incentive schemes should be more pronounced as the true probability gets further from 0.5. However, when the true probability equals 0 and 1, both BSR and QSR predict that risk-averse subjects will choose the true probability as their reported probability. Thus, the difference between the BSR and the QSR will be small for probabilities near 0.5 and larger for more extreme probabilities, but getting smaller again as the true probability gets to 0 or 1. Exactly for which probabilities the QSR starts reporting predictions close to true probability will depend on the level of risk-aversion of a subject. On the other hand, for risk-neutral agents the two schemes should lead to truthful reporting independent of the true probability. Finally, the difference between the schemes is likely to be largest for true probability close to 0.5 and miniscule for true probabilities close to 0 or 1 when the agent is risk-loving. In Table 3, we control for the distance of the true probability from 0.5. In column (1), which includes all subjects, coefficient of the dummy BSR is statistically insignificant. This suggests that there is no significant difference between the performances of the two schemes when the objective probability is around 0.5. Coefficient of the interaction term of BSR and the distance of the true probability from 0.5 is positive and significant at the 10% level. Thus, the further the true probability is from 0.5, the better the BSR performs than the QSR. This also holds true when we include only risk-averse subjects in column (2). We can use these coefficients to estimate the net impact of the BSR over the QSR for different levels of the objective probability. While the difference between the two schemes is not significant for probabilities relatively close to 0.5, the results suggest that the BSR outperforms the QSR at 5% significance level for more extreme probabilities. For example, we present the estimated effects of the BSR for when the true probability is 0.15 or 0.85 and 0.1 or 0.9.

[17] We classify subjects as risk-averse, risk-neutral, or risk-loving by constructing their average risk aversion coefficient, under a constant relative risk aversion (CRRA) model, using their choices of certainty equivalents of lotteries in the risk-preference elicitation round. Four subjects reported certainty equivalents that cannot be explained by expected utility theory. We have 269 observations from these 147 subjects. Among them, 108, 11, and 28 are risk-averse, risk-neutral, and risk-loving subjects, respectively. Our empirical results do not change qualitatively if we classify subjects without assuming any specific preference structure.

In these cases, BSR outperforms QSR for both the entire sample and only the risk-averse subjects. On the other hand, we find no difference between the two schemes for risk-neutral subjects for any level of true probability. Column (4) shows that, the coefficient of BSR is positive but significant only at 15% level for risk-loving subjects. Moreover, consistent with the theoretical prediction, there is no difference in the performance of the two incentive schemes for extreme probabilities.

While there is not much difference in the performances of the BSR and the QSR for risk-neutral subjects, the BSR performs better for risk-averse subjects. Moreover, this benefit of the BSR is more pronounced when true probabilities are relatively extreme. These results are in line with the theoretical predictions. Nevertheless, there is some randomness associated with the reported probabilities that results into subjects choosing probabilities that are different from the true probabilities under the BSR even though the BSR elicits probabilities closer to the true one compared to the QSR. The subjects report the true probability as the prediction only in handful of cases. Results not reported in the tables show that the shares of reported probabilities that coincide with the true probability are virtually the same under the BSR and the QSR at 14.4% and 13.7%, respectively. If we compare the NSD under the two schemes for observations where the reported probability was different from the true probability, we find that the BSR reduces NSD by 67.3 units.[18] Next, in Table 4, we analyze how the subjects choose the reported probability under the two schemes. We regress the reported probability on dummies for the two schemes (denoted by BSR and QSR) and the true probability interacted with BSR and QSR. The four columns use all subjects and only risk-neutral, risk averse, and risk-loving subjects, respectively. Column (1) shows that the reported probability puts higher weight on the true probability under the BSR than under the QSR. The other columns show that this difference come basically from the risk-averse subjects. As a result, the BSR clearly outperforms the QSR when subjects are risk-averse.

3.2 M-Experiment

In the M-experiment, subjects predicted the realized earning per share (EPS) of a stock. A total of 61 subjects participated in two sessions, each consisting of 40 periods. A subject was endowed with a stock of a new (fictitious) company at the beginning of each period. Then they learned 10 independent forecasts of the EPS and the average of these forecasts. The true EPS was drawn from a normal distribution with mean and variance of 60 and

[18] The effect is statistically significant at the 5% level.

400, respectively. Each forecast equaled the true EPS plus an error term independently drawn from a normal distribution with mean 0 and variance 8. In each period, a new stock (with a different, independently drawn, EPS) was presented.

Suppose, in a given period, a subject predicts the EPS to be M while the realized value is T. Then the squared error from the prediction equals $(T-M)^2$. Under the BSR, if the squared error was below some number K then the subject won a fixed prize of 80 points and won nothing otherwise. In each period, a new error bound K was generated from a uniform distribution on $[0,6]$.[19] The subject was informed of the realized values of K and the EPS (T) for a given period at the end of that period. As the underlying distributions are symmetric, Theorem 2 implies that any subject's optimal strategy under the BSR is to choose the value M that minimizes the expected mean squared error, $E_T\left[(T-M)^2\right]$, given the signals irrespective of her risk attitude. Given that the true EPS is drawn from $N(60,400)$, the optimal prediction is $500\bar{X}/501+60/501$ where $\bar{X}$ is the average of the 10 forecasts. Therefore, one can easily approximate the *optimal action* by $\bar{X}$. We use this as the *optimal action* in the M-experiment for simplicity. Under the QSR, the subject received a payment of $90-25\,(T-M)^2$. In the periods when this value was negative for a subject, she received negative points for that period. Note that reporting the average forecast is (nearly) optimal also for weakly risk-averse expected utility maximizers according to Bhattacharya and Pfleiderer (1985). We chose the parameters for the QSR (i.e., 90 and 25) such that the mean and variance of earnings when subjects followed the *optimal action* are the same for both scoring rules.

We used the BSR in the first 20 periods and the QSR in the remaining 20 periods in session 1 and the order of the incentive schemes was switched in session 2. We used the same set of 40 stocks and corresponding forecasts in the two sessions. Our data set includes 2436 observations from 61 subjects as some observations were missing because of some unknown but likely technical reason. Subjects were paid for the total points earned in the 40 periods at the rate of 16 points = HKD 1. The average payment to a subject was above HKD 164 (USD 21). The sessions were conducted in English.

3.2.1 Results

We define NSD as the negative squared difference between the predicted value of EPS and the average of the forecasts. The last column of Table 1 show that the average NSD

[19] By calculating the distribution of the posterior mean, we estimate that the theoretical probability of the error being above 6 was about 0.6% when a subject reported a prediction that maximized her expected utility. Only 60 among the 2436 observations had squared errors larger than 6.

is very small and not significantly different under the two incentive schemes. Table 5 presents results from regressing NSD on the binary variable BSR. In column (2), we also control for the subject's experience with that particular scheme.[20] The differences between performances under the two scoring rules are not statistically significant, nor economically important. We do not find any statistically significant effect of experience on performance either. Here *Experience* is a variable that equals the period number in periods 1 to 20 and the period number minus 20 in periods 21 to 40. It denotes the number of periods the subject has experienced under that incentive scheme. The predictions are very close to the average of the 10 forecasts under both schemes.[21] Note that the QSR may not be incentive compatible if the subject is not an expected utility maximizer. Nevertheless, given how closely subjects choose predictions to the optimal action, this does not seem to be of great concern.

3.3 Summary

To summarize, we investigated subjects' behavior under both binarized and quadratic scoring rules. For the P-experiment, where risk-aversion theoretically does not lead to truthful revelation of the objective probability under the QSR but does so under the BSR, subjects take actions closer to reporting the true probability when we use the BSR. For the M-experiment, subject behavior is not significantly different under the two schemes. These results are consistent with the theoretical predictions from this paper and Bhattacharya and Pfleiderer (1985). Thus, we provide a simple scoring rule that is, in theory, more general and superior in eliciting beliefs and show that this rule indeed performs better than the widely-used quadratic scoring rule in standard experiments of belief elicitation.

Our results are in stark contrast to the results of Selten, Sadrieh, and Abbink (1999) who found that a lottery payoff rule does not work well compared to deterministic scoring rules. Nevertheless, we may be able to interpret our results using the background risk hypothesis that they propose to explain their findings. The background risk hypothesis can be described in the following way: When an agent faces a high risk in the income, she may become more risk sensitive. A risk-sensitive agent may be more "capricious" in the sense that she changes her behavior sometimes toward the direction of risk-aversion

[20] We also consider specifications that include individual fixed effects or time fixed effects. However, the results are almost identical to those reported here and are omitted.

[21] Regressing the reported prediction on the average, maximum, minimum and the standard deviations of the 10 forecasts, we find that the weight on the average of forecasts is not statistically different from 1 and the weights on other variables are not statistically significant.

but also sometimes toward the direction of risk-taking. Therefore, when the risk of the income is high, she may deviate more from the prediction of the expected utility theory. In their experiments, the variance of income under the lottery payoff scheme is much higher than that under the monetary payoff scheme. On the other hand, in our M-experiment, we set the mean and the variance of the payoff such that those are the same under both schemes if an agent behaves optimally. Therefore, the background risk hypothesis does not indicate that subjects should behave more erratically under the BSR in the M-experiment. For the P-experiment, while we choose the parameters so that the mean of the income under the two schemes are the same, the variances are different. The variance under the BSR is generally higher than that under the QSR. However, simple calculations show that the difference is the largest when the true probability is 50%. The difference becomes small as the true probability approaches 0 or 1. This fact may explain the finding that while the QSR seems to work well when the true probability is around 0.50, the BSR can elicit the true probability better when the true probability is away from 0.50 where the background risks of the BSR and the QSR are similar.

4 Conclusions

This paper introduces a general mechanism to create incentive compatible scoring rules for eliciting an economic agent's beliefs of without making any strong assumption on her risk preference. We also show that the resulting scoring rules work even under some non-expected utility theory frameworks. We observe that simple scoring rules based on this mechanism perform better than the quadratic scoring rule in laboratory experiments. Given that our scoring rules are theoretically superior to commonly used scoring rules, the mechanism can be used to generate appropriate scoring rules for more sophisticated settings. For example, in a companion paper, Hossain and Okui (2011) investigate how people analyze independent and correlated signals about a random variable. Specifically, they test how well people can differentiate relative importance of independent and correlated signals. While subjects choose their prediction in an overall unbiased way, they put sub-optimally high weight on the correlated signals. They also utilize the binarized scoring rule to measure overconfidence. In general, as maximization of expected payoff and expected utility are equivalent under the binarized scoring rule, it can be used in determining payoffs in experimental games to circumvent the issue of agents not being risk-neutral.

A Appendix

Proof of Theorem 1

Proof. The agent prefers a lottery with a higher probability of receiving reward A, denoted by $P(\theta)$ in the context of the BSR. Thus, her optimization problem is equivalent to maximizing $P(\theta)$. It follows that

$$\arg\max_{\theta\in\Theta} V(L(P(\theta))) = \arg\max_{\theta\in\Theta} P(\theta).$$

Since

$$P(\theta) = E_K E_X\left[\mathbf{1}_{\{l(\theta,X)<K\}}\right] = E_X\left[1 - \frac{1}{\overline{K}} l(X,\theta)\right] = 1 - \frac{1}{\overline{K}} E_X\left[l(X,\theta)\right],$$

it holds that

$$\arg\max_{\theta\in\Theta} P(\theta) = \arg\min_{\theta\in\Theta} E_X\left[l(X,\theta)\right].$$

Therefore, the θ that minimizes the expected loss also maximizes the agent's preference functional. □

Proof of Theorem 2

Proof. Maximizing V is equivalent to maximizing the probability of winning A. This probability is[22]

$$E_X\left[\mathbf{1}_{\{(X-\theta)^2\leq\overline{K}\}}\left(1 - \frac{1}{\overline{K}}(X-\theta)^2\right)\right] = \int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}}\left(1 - \frac{1}{\overline{K}}(X-\theta)^2\right) f(X)dX.$$

The first order condition of this maximization problem is

$$-\int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} \frac{2}{\overline{K}}(\theta - X) f(X)dX = 0,$$

by the Leibniz rule. The solution to the first order condition is

$$\theta = \left(\int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} f(X)dX\right)^{-1} \int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} X f(X)dX.$$

Since $f(X)$ is a density function, it holds that

$$\left(\int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} f(X)dX\right)^{-1} \to 1$$

[22] Note that when the loss function is bounded, we take $\bar{K}$ such that $\mathbf{1}_{\{l(X,\theta)\leq\bar{K}\}} = 1$. The indicator function therefore does not appear in the proof of Theorem 1.

as $\overline{K} \to \infty$. Next, we consider the numerator. We observe that

$$E_X[X] - \int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} Xf(X)dX = \int_{\theta+\sqrt{\overline{K}}}^{\infty} Xf(X)dX + \int_{-\infty}^{\theta-\sqrt{\overline{K}}} Xf(X)dX.$$

Since $|a|^{2+\delta}f(a) \to 0$, it holds that $|a|^{2+\delta}f(a) < \epsilon$ for any ϵ for a large enough such that $af(a) < \epsilon a^{-1-\delta}$. It therefore follows that

$$\int_{\theta+\sqrt{\overline{K}}}^{\infty} Xf(X)dX < \epsilon \int_{\theta+\sqrt{\overline{K}}}^{\infty} X^{-1-\delta}dX = \epsilon\frac{1}{\delta}(\theta + \sqrt{\overline{K}})^{-\delta}$$

for $\overline{K}$ large enough. The term $(\theta+\sqrt{\overline{K}})^{-\delta}$ can be made arbitrarily small by taking $\overline{K}$ large enough. This shows that $\int_{\theta+\sqrt{\overline{K}}}^{\infty} Xf(X)dX \to 0$. Similarly, we have $\int_{-\infty}^{\theta-\sqrt{\overline{K}}} Xf(X)dX \to 0$. These imply that

$$\left(\int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} f(X)dX\right)^{-1} \int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} Xf(X)dX \to E_X[X].$$

Note that, when $f(X)$ is symmetric around the mean, the solution to the first order condition,

$$-\int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} \frac{2}{\overline{K}}(\theta - X)f(X)dX = 0,$$

exactly equals $E_X[X]$ even for finite $\overline{K}$.

We verify that the solution to the first order condition is indeed the maximizer by checking the second order condition:

$$\begin{aligned}
& -\frac{2}{\overline{K}}(\theta - \theta - \sqrt{\overline{K}})f(\theta + \sqrt{\overline{K}}) + \frac{2}{\overline{K}}(\theta - \theta + \sqrt{\overline{K}})f(\theta - \sqrt{\overline{K}}) \\
& -\frac{2}{\overline{K}}\int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} f(X)dX \\
= \; & \frac{2}{\overline{K}}\left(\sqrt{\overline{K}}f(\theta + \sqrt{\overline{K}}) + \sqrt{\overline{K}}f(\theta - \sqrt{\overline{K}}) - \int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} f(X)dX\right),
\end{aligned}$$

by the Leibniz rule. The term $\sqrt{\overline{K}}f(\theta + \sqrt{\overline{K}}) + \sqrt{\overline{K}}f(\theta - \sqrt{\overline{K}})$ can be made arbitrarily small by taking $\overline{K}$ large enough by the assumption that $|a|^{2+\delta}f(a) \to 0$ as $a \to \infty$ and as $a \to -\infty$. The term $\int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} f(X)dX$ can be made arbitrarily close to 1 by taking $\overline{K}$ large enough. Therefore, the second order condition is satisfied for large $\overline{K}$. □

References

[1] Allen, Franklin (1987): "Discovering Personal Probabilities When Utility Functions are Unknown," *Management Science*, 33(4), 542-544.

[2] Andersen, Steffen, John Fountain, Glenn W. Harrison and E. Elisabet Rutström (2010): "Estimating Subjective Probabilities," Working paper, 2010-06, Center for the Economic Analysis of Risk, Georgia State University.

[3] Andersen, Steffen, John Fountain, Glenn W. Harrison, Arne Risa Hole and E. Elisabet Rutström (2012): "Inferring Beliefs as Subjectively Imprecise Probabilities," *Theory and Decision*, 73(1), 161-184.

[4] Arellano, Manuel (1987), "Computing Robust Standard Errors for Within-groups Estimators," *Oxford Bulletin of Economics and Statistics*, 49(4), 431-435.

[5] Becker, Gordon M., Morris H. Degroot and Jacob Marschak (1964): "Measuring Utility by a Single-response Sequential Method," *Behavioral Science*, 9(3), 226-232.

[6] Berg, Joyce E., Lane A. Daley, John W. Dickhaut, and John R. O' Brien (1986): "Controlling Preferences for Lotteries on Units of Experimental Exchange," *Quarterly Journal of Economics*, 101(2), 281-306.

[7] Bhattacharya, Sudipto and Paul Pfleiderer (1985): "Delegated Portfolio Management," *Journal of Economic Theory*, 36(1), 1-25.

[8] Blanco, Mariana, Dirk Engelmann, Alexander K. Koch and Hans-Theo Normann (2010): "Belief Elicitation in Experiments: Is There a Hedging Problem?," *Experimental Economics*, 13, 412-438.

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

[10] Chew, Soo Hong and Jacob S. Sagi (2006): "Event Exchangeability: Probabilistic Sophistication Without Continuity or Monotonicity," *Econometrica*, 74(3), 771-786.

[11] De Finetti, Bruno (1974): *Theory of Probability*, Vol. 1, New York: Wiley.

[12] Diamond, Peter A. (1967): "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment," *Journal of Political Economy*, 75(6), 765-766.

[13] Fischbacher, Urs (2007): "z-Tree - Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, 10(2), 171-178.

[14] Gneiting, Tilmann and Adrian E. Raftery (2007): "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102(477), 359-378.

[15] Grant, Simon (1995): "Subjective Probability without Monotonicity: Or how Machina's Mom May also Be Probabilistically Sophisticated," *Econometrica*, 63(1), 159-191.

[16] Grether, David M. (1980): "Bayes Rule as a Descriptive Model: The Representative Heuristic," *Quarterly Journal of Economics*, November, 537-557.

[17] Grether, David M. (1981): "Financial Incentive Effects and Individual Decision making," Social Science Working Paper 401, California Institute of Technology.

[18] Grether, David M. (1992): "Testing Bayes Rule and the Representative Heuristic: Experimental Evidence," *Journal of Economic Behavior and Organization*, 17, 31-57.

[19] Hao, Li and Daniel Houser (2012): "Belief Elicitation in the Presence of Naïve Respondents: An Experimental Study," *Journal of Risk and Uncertainty*, 44(2), 161-180.

[20] Harrison, Glenn W., Jimmy Martínez-Correa, and J. Todd Swarthout (2012): "Inducing Risk Neutral Preferences with Binary Lotteries: A Reconsideration," Working Paper, 2012-02, Center for the Economic Analysis of Risk, Georgia State University.

[21] Hollard, Guillaume, Sebastien Massoni, and Jean-Christophe Vergnaud (2010): "Subjective Beliefs Formation and Elicitation Rules: Experimental Evidence," Working Paper, Universite Paris-1.

[22] Holt, Charles A. (1986): "Preference Reversals and the Independence Axiom," *American Economic Review*, 76(3), 508-515.

[23] Holt, Charles (2007): *Markets, Games & Strategic Behavior*, Boston, Pearson/Addison-Wesley.

[24] Holt, Charles A. and Angela M. Smith (2009): "An Update on Bayesian Updating," *Journal of Economic Behavior & Organization*, 69, 125-134.

[25] Hossain, Tanjim and Ryo Okui (2011): "Information Aggregation in a Laboratory Financial Market," working paper.

[26] Hurley, Terrance M., Nathanial Peterson, and Jason F. Shogren (2007): "Belief Elicitation: An Experimental Comparison of Scoring rule and Prediction Methods," Working paper, University of Minnesota.

[27] Hurley, Terrance M. and Jason F. Shogren (2005): "An Experimental Comparison of Induced and Elicited Beliefs," *Journal of Risk and Uncertainty*, 30(2), 169-188.

[28] Kadane, Joseph B. and Robert L. Winkler (1988): "Separating Probability Elicitation from Utility," *Journal of the American Statistical Association*, 88(402), 357-363.

[29] Karni, Edi (2009): "A Mechanism for Eliciting Probabilities," *Econometrica*, 77(2), 603-606.

[30] Karni, Edi, and Zvi Safra (1987): ""Preference Reversal" and the Observability of Preferences by Experimental Methods," *Econometrica*, 55(3), 675-685.

[31] Karni, Edi and Zvi Safra (1995): "The Impossibility of Experimental Elicitation of Subjective Probabilities," *Theory and Decision*, 38(3), 313-320.

[32] Laury, Susan K., Melayne Morgan McInnes, and J. Todd Swarthout (2012): "Avoiding the Curves: Direct Elicitation of Time Preferences," *Journal of Risk and Uncertainty*, 44(3), 181-217.

[33] Machina, Mark J. (1982): ""Expected Utility" Analysis without the Independence Axiom," *Econometrica*, 50(2), 277-323.

[34] Machina, Mark J. (1989): "Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty," *Journal of Economic Literature*, 27(4), 1622-1668.

[35] Machina, Mark J. and David Schmeidler (1992): "A More Robust Definition of Subjective Probability, " *Econometrica*, 60(4), 745-780.

[36] McKelvey, Richard D. and Talbot Page (1990): "Public and Private Information: An Experimental Study of Information Pooling," *Econometrica*, 58(6), 1321-1339.

[37] Möbius, Markus M., Muriel Niederle, Paul Niehaus and Tanya Rosenblat (2007): "Gender Differences in Incorporating Performance Feedback," Working Paper, Harvard University, Cambridge.

[38] Offerman, Theo, Joep Sonnemans, Gijs van de Kuilen and Peter P. Wakker (2009): "A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes," *Review of Economic Studies*, 76, 1461-1489.

[39] Qu, Xiangyu (2012): "A Mechanism for Eliciting a Probability Distribution," *Economics Letters*, 115, 399-400.

[40] Quiggin, John (1982): "A Theory of Anticipated Utility," *Journal of Economic Behavior and Organization*, 3, 323-343.

[41] Roth, Alvin E. and Michael W. K. Malouf (1979): "Game-Theoretic Models and the Role of Information in Bargaining," *Psychological Review*, 86(6), 574-594.

[42] Savage, Leonard J. (1971): "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association*, 66(336), 783-801.

[43] Schlag, Karl H. and Joël van der Weele (2009): "Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming risk-neutrality," Working Paper, Universitat Pompeu Fabra, Barcelona.

[44] Selten, Reinhard, Abdolkarim Sadrieh, and Klaus Abbink (1999): "Money Does Not Induce risk-neutral Behavior, But Binary Lotteries Do Even Worse," *Theory and Decision*, 46(3), 211-249.

[45] Smith, Cedric A. B. (1961): "Consistency in Statistical Inference and Decision," *Journal of the Royal Statistical Society*, Series B, 23(1), 1-25.

[46] Trautmann, Stefan T. and Gijs van de Kuilen (2011): "Belief Elicitation: A Horse Race among Truth Serums," Working Paper, Tilburg University.

	P-experiment	M-experiment
Average of NSD	-138.203	-0.218
Standard Deviation	279.145	0.584
Observations	276	2436
Average of NSD under the BSR	-105.891	-0.207
Standard deviation	175.994	0.523
Observations	137	1216
Average of NSD under the QSR	-170.050	-0.230
Standard Deviation	350.359	0.639
Observations	139	1220
t-test	2.021	1.019
p-value	0.044	0.308

Note: NSD is the negative of the square of the difference between the reported number and the action that minimizes the expected loss (it is the true probability in percentage in the P-experiment and the average of forecasts in the M-experiment). t-test is the value of the t-test statistic for the null hypothesis that the mean of the NSD under the BSR is equal to that under the QSR. Heteroskedasticity is allowed and the p-value is computed by its asymptotic distribution.

Table 1: Summary of NSD from both experiments

Dependent Variable	NSD	NSD	NSD	NSD
Constant	-170.05***	-200.52***	-98.50	-99.96**
	(29.72)	(39.18)	(66.23)	(39.11)
BSR	64.16**	80.08**	27.00	25.24
	(29.69)	(39.64)	(26.29)	(41.87)
Adjusted R^2	0.013	0.012	-0.050	-0.014
Sample	Full	Risk-averse	Risk-neutral	Risk-loving
# of Observations	276	198	20	51

Note: "***" and "**" indicate significance at the 1% and 5% levels, respectively. Heteroskedasticity and autocorrelation robust standard errors are in parentheses. NSD is the negative of the square of the difference between the reported prediction and the true probability (in percentage terms); BSR is a dummy variable equaling 1 if the BSR is used.

Table 2: Subject performance under the BSR and the QSR, P-experiment.

Dependent Variable	NSD	NSD	NSD	NSD
Constant	33.79	93.68	-3.41	-160.03**
	(66.85)	(100.24)	(50.77)	(61.49)
BSR	-95.28	-176.41*	97.47	102.80
	(73.18)	(106.16)	(133.20)	(69.96)
$\|Trueprob - 50\|$	-6.85***	-9.70**	-3.80	1.99
	(2.92)	(4.18)	(2.94)	(1.53)
$\|Trueprob - 50\| \times BSR$	5.27*	8.38*	-2.37	-2.66
	(3.12)	(4.36)	(5.04)	(1.89)
Adjusted R^2	0.047	0.065	-0.063	-0.044
The impact of the BSR when true probability is 0.15 or 0.85				
$\beta_{BSR} + 35\beta_{\|Trueprob-50\|\times BSR}$	89.10**	117.06**	14.38	9.70
	(43.27)	(56.24)	(56.84)	(44.24)
when true probability is 0.10 or 0.90				
$\beta_{BSR} + 40\beta_{\|Trueprob-50\|\times BSR}$	115.44**	158.98**	2.51	-3.60
	(57.61)	(76.27)	(78.94)	(47.47)
Sample	Full	Risk-averse	Risk-neutral	Risk-loving
# of Observations	276	198	20	51

Note: "***", "**", and "*" indicate significance at the 1%, 5%, and 10% levels, respectively. Heteroskedasticity and autocorrelation robust standard errors are in parentheses. NSD is the negative of the square of the difference between the reported prediction and the true probability (in percentage terms); BSR is a dummy variable equaling 1 if the BSR is used; $Trueprob$ is the true probability in percentage terms.

Table 3: Subject performance under the BSR and the QSR in relation to the difference between the true probability and 0.50, P-experiment.

Dependent Variable	*Reported*	*Reported*	*Reported*	*Reported*
BSR	2.48	2.63	3.52	1.78
	(1.53)	(1.89)	(2.26)	(3.40)
$BSR \times Trueprob$	0.92^{***}	0.91^{***}	0.87^{***}	0.97^{***}
	(0.03)	(0.03)	(0.08)	(0.07)
QSR	8.33^{***}	9.65^{***}	14.23	2.85
	(2.09)	(2.51)	(10.26)	(3.51)
$QSR \times Trueprob$	0.84^{***}	0.80^{***}	0.78^{***}	0.99^{***}
	(0.04)	(0.04)	(0.14)	(0.06)
Adjusted R^2	0.859	0.841	0.887	0.907
Wald test for the equivalence between the behaviors				
$H_0 : \beta_{BSR} = \beta_{QSR}$	7.17	7.24	1.60	0.98
and $\beta_{BSR \times Trueprob} = \beta_{QSR \times Trueprob}$	(0.028)	(0.027)	(0.450)	(0.611)
Sample	Full	Risk-averse	Risk-neutral	Risk-loving
# of Observations	276	198	20	51

Note: "***" indicates significance at the 1% level. Heteroskedasticity and autocorrelation robust standard errors are in parentheses under coefficient estimates. *Reported* is the reported prediction in percentage terms; BSR is a dummy variable equaling 1 if the BSR is used; $Trueprob$ is the true probability in percentage terms; QSR is a dummy variable equaling 1 if the QSR is used. In parentheses under the values of the test statistics are p-values.

Table 4: Subject behaviors in the P-experiment.

Dependent Variable	NSD	NSD
Constant	-0.230***	-0.294***
	(0.026)	(0.055)
BSR	0.023	0.018
	(0.026)	(0.072)
$Experience$		0.006
		(0.004)
$BSR \times Experience$		0.000
		(0.005)
Adjusted R^2	0.000	0.003
Wald test for the impact of $Experience$		
H_0: $\beta_{Experience} = 0$ and $\beta_{BSR \times Experience} = 0$		4.806
		(0.090)
Wald test for the impact of BSR		
H_0: $\beta_{BSR} = 0$ and $\beta_{BSR \times Experience} = 0$		1.144
		(0.564)
The effect of $EXPERIENCE$ under the BSR:		
$\beta_{EXPERIENCE} + \beta_{BSR \times EXPERIENCE}$		0.006*
		(0.039)
Observations	2436	2436

Note: "***" and "* indicate significance at the 1% and 10% levels, respectively. Heteroskedasticity and autocorrelation robust standard errors are in parentheses under coefficient estimates. In parentheses under the values of the test statistics are p-values. NSD is the negative of the square of the difference between the reported number and the average forecast; BSR is a dummy variable equaling 1 if the BSR is used; The variable $Experience$ takes the value of 1 to 20 to denote the period number under the particular incentive scheme.

Table 5: Subject performance under the BSR and the QSR, M-experiment