The Binarized Scoring Rule of Belief Elicitation[*]

Tanjim Hossain
University of Toronto

Ryo Okui
Kyoto University

The First Version: December 2009
This Version: October 2010

Abstract

We introduce a simple method of constructing an incentive compatible scoring rule to elicit an agent's belief about a random variable. The method does not depend on the exact form of the agent's utility function. Independent of her risk-preference, it is optimal for her to report the value that minimizes the expected value of a loss function specified in the incentive scheme. Under this incentive scheme, the agent receives a fixed prize when her prediction error, defined by the relevant loss function, is smaller than a randomly generated value 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. Our experimental results suggest that this scoring rule performs better than the quadratic scoring rule.

Keywords: Scoring rule, subjective belief, randomization, prediction, laboratory experiment.

JEL classification: C91, D81, D83, D84.

[*] We thank Robin Chark, Eric Set, and Keigo Inukai for their excellent research assistance, Phil Dybvig, Haluk Ergin, Glenn Harrison, Atsushi Kajii, Sachar Kariv, Nicolas Lambert, David Levine, Colin Stewart, Norio Takeoka, and seminar participants at the University of Washington St. Louis and the spring meeting of the Japanese Economic Association for helpful comments and Sudipto Dasgupta for initially getting us interested in the problem. Yoichi Hizen 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 of HKUST under Project No. RPC06/07.BM04. Please direct all correspondence to Tanjim Hossain at tanjim.hossain@utoronto.ca.

1 Introduction

We introduce a parsimonious but extremely general way of constructing a scoring rule to elicit an agent's beliefs about a random variable. Using this method, an expected utility maximizing agent reports the value that minimizes the expected value (conditional on her information) of a loss function specified in the incentive scheme, 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 a number generated independently of the random variable in question and the agent's reported value. Specifically, after observing information about a random variable, the agent reports her prediction of some feature of that variable which may incorporate her beliefs. Then, she receives a fixed prize if her prediction error, defined by the relevant loss function, is smaller than a randomly generated number from a uniform distribution and earns a smaller prize or a penalty otherwise. In other words, she receives the prize with a probability which is determined by the value of the loss function. An expected utility maximizing agent's objective, thus, becomes maximizing the probability of winning the reward and this probability is proportional to the expected loss. Under relatively weak assumptions, her optimal action is to minimize the expected value of the loss function. We term this incentive scheme as the Binarized Scoring Rule (BSR).

Eliciting an economic agent's beliefs concerning a probabilistic event (which may be objectively specified) is an important problem. Many have suggested scoring rules to incentivize truthful communication of agent's beliefs. Many such mechanisms, such as the proper scoring rules suggested by Savage (1971) or the methods of promissory notes by De Finetti (1974), however, make agents report the probability under their belief only under risk-neutrality. For example, we may want to know an agent's subjective probability of the occurrence of an event. The agent has no incentive to report her subjective probability if we simply ask her to do so. A well-known method for providing the agent with monetary incentives to truthfully report her subjective probability is the quadratic scoring rule (QSR) of Brier (1950). Suppose the agent is paid $a - b(1-p)^2$ if the event occurs and $a - bp^2$ if the event does not occur where p is her reported number. 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 than reporting her true belief. Generally speaking, under risk-aversion, the marginal utility of the monetary payment to the agent confounds her beliefs. As this example shows, it is not an easy task to elicit an agent's belief without making an assumption on her utility structure.

As discussed in the beginning of the paper, this paper proposes an innovative way to construct a scoring rule to elicit some feature of the belief without assuming risk neutrality that can be applied in a wide array of situations.[1] 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 amount of money if the score is less than a uniform random variable. For example, to elicit the subjective probability of an event, we binarize the QSR such that the agent receives a fixed reward when $(1-p)^2 < k$, where k is the realized value of a uniform random variable, if the event occurs and if the event does not occur, the reward is given when $p^2 < k$. Our method induces risk-neutral behavior by agents even if they are risk-averse or risk-loving by basically turning the incentive scheme into a binary lottery. We can adjust the loss function according to our goal for eliciting the agent's be-

[1] Alternative approaches are to recover true belief from the action of the agent that potentially is different from the true belief because of risk aversion or some other reason. Andersen, Fountain, Harrison, and Rutström (2010) jointly estimate the risk attitude and the true subjective probability. Offerman, Sonnemans, van de Kuilen, and Wakker (2009) provide a way to correct the reported probability to recover the subjective probability.

liefs. For example, an expected utility maximizing subject will minimize the mean squared error of her estimate of the expected value of a scalar random variable if we use a quadratic loss function. Moreover, by choosing the loss function appropriately, we can elicit the agent's belief about the median of the variable or some other characteristic of the variable. Thus, we do not just provide a scoring rule, rather we suggest a method to device appropriate scoring rules according to the belief elicitation objective. The binarized scoring rule can even be used when the agent's decision mechanism is described not by expected utility but by Machina's (1982) utility function or prospect theory suggested by Kahneman and Tversky (1979).

Several previous studies developed scoring rules that do not require assumptions on utility function. For eliciting the mean of the subjective probability distribution, 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 (weakly) risk-averse. For eliciting the probability of a certain event, Allen (1987) proposed a randomized scoring rule for a very specific scenario where the agent receives a fixed reward using a lottery when she has to estimate the probability of a certain event (of only two possible events) happening. When we elicit the probability of an event, then our BSR becomes equivalent to this method (see Schlag and van der Weele, 2009). Recently, Karni (2009) proposed a more complicated mechanism that involves two layers of randomization.[2] His method is to give one of two different lotteries depending on the probability reported by the agent and the realization of a random number chosen by the experimenter. This mechanism is specifically designed for eliciting a probability and it is difficult to extend the idea to other settings. On the other hand, our simple approach can be used in eliciting almost any feature of the

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

subjective belief as long as there exists a scoring rule that induces that feature under risk-neutrality.[3] For example, Friedman (1983) developed scoring rules for eliciting probability distribution that requires risk-neutrality. Using Friedman's score as the loss function in BSR, we can even elicit a probability distribution function without assuming the form of the agent's utility function.

Independently of our work, Schlag and van der Weele (2009) have developed a randomized scoring rule for eliciting probability and other related parameters. Their method, which is an extension of Allen's, is essentially same as ours, but can be applied to a limited set of situations. We emphasize the generality of the proposed method and present theorems that justify the use of our BSR under a very general framework including cases in which the loss function is unbounded. The validity of the BSR under non-expected utility theory frameworks suggested by Machina (1982) and Kahneman and Tversky (1979) is a new result too. We also present experimental results that support the use of the BSR. On the other hand, Schlag, and van der Weele (2009) emphasize the relationship between the fixed prize method of BSR (and Allen, 1987) and the methods of Grether (1981), Holt (2007) and Karni (2009). Theoretically, our method formalizes and generalizes the intuition in Roth and Malouf (1979) who suggest paying subjects with tickets for a binary lottery instead of direct monetary payments to induce risk-neutral behavior.[4] However, Selten, Sadrieh, and Abbink (1999) contend that, rewards

[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 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 problem 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] 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.

determined by a lottery do not induce risk-neutral behavior in subjects in a laboratory experiment. In fact, they performed much worse than money prizes in their experiment.

Since our scoring rule involves randomization of the reward, it is important to check whether our scoring rule suffers from the problem reported by them. We run experiments using our scoring rule and the quadratic scoring rule. The purposes of running the experiments are to illustrate how the binarized scoring rule can be utilized to incentivize truthful reporting of beliefs by subjects and to compare the performances of our scoring rule to that of the quadratic scoring rule. There are two types of experiments. One of them 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 and we investigate whether subjects report the "correct" probability or mean under these two scoring rules. In the experiment where subjects report a probability, their predictions are closer to the objective probability under the BSR than under the quadratic scoring rule implying that the BSR performs better in this setting. Moreover, we find that our scoring rule performs as well as the quadratic scoring rule in the experiment in which we elicit the mean as is theoretically predicted. These results are in contrast with those of Selten, Sadrieh, and Abbink (1999).

There has been a number of experimental investigations to elicit belief of agents in the laboratory.[5] Some of them use deterministic scoring rules (such as the quadratic scoring rule) and some of them use probabilistic scoring rules like ours. Nevertheless, there are not too many previous studies that compare the performances of various scoring rules apart from Andersen, Fountain,

[5] See, for example, 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) cite papers that use scoring rules in various different fields.

Harrison, and Rutström (2010) who compare the QSR and a linear scoring rule.[6] Thus, reporting the results of the experiments that compare various scoring rules is another contribution of this paper to the literature.

The rest of the paper is structured as follows. Section 2 presents the binarized scoring rule and the theorems that justify the use of the scoring rule. Section 3 describes the design of the laboratory experiment and discusses the results from these experiment. Section 4 concludes.

2 The Binarized Scoring Rule

This section presents the *Binarized Scoring Rule* or *BSR* to elicit beliefs. We present a theorem that justifies the use of the proposed scoring rule and demonstrate that the scoring rule works under general situations. First, we explain the setting. An expected utility maximizing economic agent has information that can be summarized by a σ-field $\mathcal{F}$. Let X be a random variable that is $\mathcal{F}$-measurable. We do not impose any other restriction on X. The random variable X can be anything so that it may be a scalar, vector or even an infinite-dimensional stochastic process. Let $\theta \in \Theta$ be the action taken by the agent, or what she reports about the variable X after observing the information represented by $\mathcal{F}$. The dimension of θ is not restricted to one. It can even be a vector or a function. Let $l(\theta, X)$ be a scalar valued function of θ and X. We call the function l the loss function. Our goal is to devise an incentive structure or scoring rule under which the agent will report a value of θ that minimizes $E(l(\theta, X)|\mathcal{F})$ irrespective of the exact form of her utility function. Here the expectation is taken with respect to the agent's belief conditional on $\mathcal{F}$. As an example, consider the situation in which X is a scalar and we want to know the expected value of X under the agent's

[6] 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).

belief. In this case, we use $l(\theta - X) = (\theta - X)^2$. Then, $E(X|\mathcal{F})$ minimizes $E(l(\theta, X)|\mathcal{F})$. Another example is when we want to elicit the median of a scalar X. The appropriate loss function for this case is $l(\theta - X) = |\theta - X|$. A non-trivial question that is addressed in this section is how to incentivize the agent to report truthfully the value of θ that minimizes $E(l(\theta, X)|\mathcal{F})$. When the agent is risk-neutral, then paying the agent the amount $-l(\theta, X)$ after X is realized provides a sufficient incentive. However, the value of θ that maximizes $E(u(-l(\theta, X))|\mathcal{F})$, where u is the agent's utility function, is in general different from the value of θ that minimizes $E(l(\theta, X)|\mathcal{F})$. This is particularly problematic if we do not know the form of the utility function.

To overcome this problem, we consider the following scoring rule. Let K be a uniform random variable whose support is $[0, \overline{K}]$ (i.e., $K \sim U[0, \overline{K}]$), where $\overline{K}$ is a positive number. The agent receives a reward A if $l(X, \theta) < K$ and B if $l(X, \theta) \geq K$ where she strictly prefers receiving A over receiving B. Let us define $\mathbf{1}_{\{l(X,\theta)<K\}}$ to be the event where the loss function $l(X, \theta)$ is below K. This scheme provides an incentive to the agent to report the value of θ that minimizes $E(l(\theta, X)|\mathcal{F})$ as proved below.

To summarize, the time-line of the mechanism is as follows:

1. The agent receives information represented by a σ-field $\mathcal{F}$.

2. The agent reports θ to the experimenter.

3. X is realized.

4. The experimenter draws K from $U[0, \overline{K}]$ independent of the realization of X or the reported θ.

5. The agents receives A if if $l(X, \theta) < K$ and B if $l(X, \theta) \geq K$.

The following assumption states the set of conditions needed for this mechanism to work.

Assumption 1. *i) The random variable X is $\mathcal{F}$-measurable.*
ii) The variable K is drawn from $U[0, \overline{K}]$ where K is independent of X.
iii) The agent values getting A more than getting B, $u(A) > u(B)$.
iv) Realized value of the loss function is non-negative $l(X, \theta) \geq 0$.
v) The expressions $\arg\min_{\theta\in\Theta} E\left(l(X,\theta)|\mathcal{F}\right)$ *and* $\arg\max_{\theta\in\Theta} E\left(u\left(A\mathbf{1}_{\{l(X,\theta)<K\}}\right)|\mathcal{F}\right)$ *are well-defined.*

The following theorem justifies the use of this mechanism. It shows that the value of θ that maximizes the expected utility under our scoring rule is the same as the minimizer of $E(l(X,\theta)|\mathcal{F})$ 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 $l(X, \theta) < \overline{K}$ for any X and θ. Then,*

$$\arg\max_{\theta\in\Theta} E\left(u\left(A\mathbf{1}_{\{l(X,\theta)<K\}}\right)|\mathcal{F}\right) = \arg\min_{\theta\in\Theta} E\left(l(X,\theta)|\mathcal{F}\right).$$

Proof. Because $u(A) > u(B)$, maximizing the expected utility is equivalent to the problem of maximizing the expected payoff or the expected value of the probability of winning the reward:

$$E\left(\mathbf{1}_{\{l(\theta,X)<K\}}|\mathcal{F}\right),$$

which equals to

$$E\left(1 - \frac{1}{\overline{K}} l(\theta, X)|\mathcal{F}\right).$$

It therefore follows that

$$\begin{aligned} \arg\max_{\theta\in\Theta} E\left(u\left(A\mathbf{1}_{\{l(X,\theta)<K\}}\right)|\mathcal{F}\right) &= \arg\max_{\theta\in\Theta} E\left(1 - \frac{1}{\overline{K}} l(\theta, X)|\mathcal{F}\right) \\ &= \arg\min_{\theta\in\Theta} E\left(l(\theta, X)|\mathcal{F}\right). \end{aligned}$$

We note that the boundedness assumption $l(X, \theta) < \overline{K}$ is needed because

$$\begin{aligned} E\left(\mathbf{1}_{\{l(\theta,X)<K\}}|\mathcal{F}\right) &= E\left(\mathbf{1}_{\{l(\theta,X)\leq\overline{K}\}}\mathbf{1}_{\{l(\theta,X)<K\}}|\mathcal{F}\right) \\ &= E\left(\mathbf{1}_{\{l(\theta,X)\leq\overline{K}\}}\left(1-\frac{1}{\overline{K}}l(\theta,X)\right)|\mathcal{F}\right) \end{aligned}$$

is not necessarily equal to $E\left(1 - l(\theta, X)/\overline{K}|\mathcal{F}\right)$ if $\mathbf{1}_{\{l(\theta,X)\leq\overline{K}\}} = 1$ is not assumed. □

The idea of the proof is the following. Because the agent receives the reward only when the loss is less than K, maximizing the expected utility becomes equivalent to maximizing the probability of the loss being less than K. Because K is uniform and the loss is always in the support of K, the probability is negatively proportional to the expected loss. Therefore, maximizing the expected utility becomes equivalent to minimizing the expected loss. The theorem demonstrates that a maximizer of the expected utility must be a minimizer of the expected loss even if there multiple maximizers of the expected utility. When there are multiple maximizers, it implies that all of them are minimizers of the expected loss.

This theorem is very general. For example, there is no restriction on the distribution of X. It can be continuous, discreet or mixture. The dimension of X is not restricted. We even allow X to be a stochastic process. We also have little restriction on θ. In most applications including the experiments we describe in the following sections, θ is a scalar. However, the theorem can be applied even when θ is a function. For example, this mechanism can be applied when we are interested in the belief (which is a probability measure) itself.[7] Lastly and perhaps most importantly, the theorem requires little assumption on the form of the utility function. We need to assume that economic agents like the reward A more than B; that is, $u(A) > u(B)$, which is hardly a controversial assumption because we can easily choose A

[7] We may employ the loss functions discussed in Friedman (1983) to elicit the probability distribution function.

and B so that it satisfies this assumption. For example, when A and B are monetary awards, then setting $A > B$ would be sufficient as long as the agent like money. We also do not need to make any assumption on the initial wealth level of the agent.

This theorem relies on are that the preference of the agent can be described with the expected utility framework and that the agent correctly evaluates the probabilities. However, the expected utility theory is not required to show the validity of the BSR. In section 2.1, we show that the binarized scoring rule is incentive compatible even under some other non-expected utility theory and prospect theory. On the other hand, "probabilistic sophistication" is needed.[8] If the agent's belief does not satisfy the requirements to be a probability measure, then the expectation $E((l(X,\theta)|\mathcal{F}))$ is not well defined so as the value of θ minimizing $E((l(X,\theta)|\mathcal{F}))$. Chambers (2008) considers the elicitation by deterministic scoring rules under the max-min utility theory with risk neutrality and characterizes what the scoring rules elicit. A similar result may be obtained for the BSR but it is beyond the scope of this paper.

The limitation of the theorem that may be important in practice is the assumption that $l(X,\theta)$ is bounded by $\overline{K}$. Although we can take $\overline{K}$ so large that this boundedness assumption is practically satisfied, we may still have some concerns. Nevertheless, under many circumstances, this boundedness assumption can be relaxed. It appears that the additional assumptions required to relax the boundedness assumption are specific to the loss function. In the experiments in this paper, we consider the case in which we are interested in minimizing the mean squared error concerning the realized value of the scalar random variable X (which can be continuous or discreet). Hence, the relevant loss function is quadratic, $l(X,\theta) = (X-\theta)^2$. We show that

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

when the distribution of X is symmetric around θ and has a light tail, reporting the mean maximizes the expected utility when the loss function is quadratic. This result is summarized in the following theorem.

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

$$\arg\max_{\theta \in \Theta} E\left(u\left(A\mathbf{1}_{\{(X-\theta)^2 < K\}}\right)|\mathcal{F}\right) \to E(X|\mathcal{F}).$$

Moreover, when $f(X|\mathcal{F})$ is symmetric around the mean, there exists a finite $\widetilde{K}$ such that for any $\overline{K} > \widetilde{K}$,

$$\arg\max_{\theta \in \Theta} E\left(u\left(A\mathbf{1}_{\{(X-\theta)^2 < K\}}\right)|\mathcal{F}\right) = E(X|\mathcal{F}).$$

Proof. The function to be maximized becomes

$$\begin{aligned} & E\left(\mathbf{1}_{\{(X-\theta)^2 \le \overline{K}\}}\left(1 - \frac{1}{\overline{K}}(X-\theta)^2\right)|\mathcal{F}\right) \\ = & \int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} \left(1 - \frac{1}{\overline{K}}(X-\theta)^2\right) f(X|\mathcal{F})dX. \end{aligned}$$

The first order condition of this maximization problem is

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

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

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

Since $f(X|\mathcal{F})$ is a density function, it holds that

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

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

$$E(X|\mathcal{F}) - \int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} Xf(X|\mathcal{F})dX = \int_{\theta+\sqrt{\overline{K}}}^{\infty} Xf(X|\mathcal{F})dX + \int_{-\infty}^{\theta-\sqrt{\overline{K}}} Xf(X|\mathcal{F})dX.$$

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

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

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

$$\int_{\theta+\sqrt{\overline{K}}}^{\infty} Xf(X|\mathcal{F})dX \to 0.$$

Similarly, we have

$$\int_{-\infty}^{\theta-\sqrt{\overline{K}}} Xf(X|\mathcal{F})dX \to 0.$$

This shows that

$$\left(\int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} f(X|\mathcal{F})dX\right)^{-1} \int_{\theta-\sqrt{\overline{K}}}^{\theta+\sqrt{\overline{K}}} Xf(X|\mathcal{F})dX \to E(X|\mathcal{F}).$$

Note that, when $f(X|\mathcal{F})$ is symmetric around the mean, the solution to the first order condition,

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

exactly equals $E(X|\mathcal{F})$ 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:

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

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

□

This result is comparable to Proposition 2 in Bhattacharya and Pfleiderer (1985) which shows that the QSR can elicit the mean of a random variable with symmetric distribution for a risk-averse agent. We show that similar result holds under the BSR even if the agent is risk-averse. More importantly, BSR is incentive compatible at the limit as $\overline{K}$ approaches infinity even if the underlying distribution is not symmetric. Thus, the BSR is more widely applicable than the QSR when we want to elicit the mean of a random variable.

2.1 Departure from the Expected Utility Framework

So far, we have assumed that the agent optimizes under the expected utility paradigm. Although this is a fundamental assumption in economic theory, many studies suggest that the expected utility theory may not describe well a real decision mechanism under uncertainty and several alternative frameworks of non-expected utility have been proposed.

In this section, we discuss whether the binarized scoring rule can be used under any non-expected utility framework. Specifically, we show that, under an additional assumption, Theorems 1 and 2 can be extended to the framework where the economic agent's decision problems is summarized by the Machina (1982) theory or the prospect theory.

We note that the payment from the BSR is represented as the distribution function

$$F(a,\theta) = \begin{cases} 1 - E\left(\mathbf{1}_{\{l(\theta,X)<K\}}\middle|\,\mathcal{F}\right) & \text{if } 0 \leq a < A, \\ 1 & \text{if } a = A. \end{cases}$$

The decision made by an agent under the BSR is to choose a lottery from the set of lotteries which is indexed by θ. We observe that

$$E\left(\mathbf{1}_{\{l(\theta,X)<K\}}\middle|\,\mathcal{F}\right) = E\left(1 - \frac{1}{K}l(\theta,X)|\mathcal{F}\right) = 1 - \frac{1}{K}E\left(l(\theta,X)|\mathcal{F}\right).$$

Let $\theta^* \in \arg\min_{\theta\in\Theta} E(l(\theta,X)|\mathcal{F})$. Then $F(\cdot,\theta^*)$ stochastically dominates $F(\cdot,\theta)$ for $\theta \notin \arg\min_{\theta\in\Theta} E(l(\theta,X)|\mathcal{F})$. It implies that the BSR works under any framework that satisfies the property that an agent prefers a lottery whose distribution is stochastically dominating to a stochastically dominated one. This property is called *monotonicity with respect to stochastic dominance* by Machina and Schmeidler (1992, p754).[9] Many non-expected utility theories satisfy this property. In particular, we consider the non-expected utility theory examine by Machina (1982).

The decision theory considered in Machina (1982) does not depend on the independence axiom, which is arguably the most controversial assumption in the expected utility theory. The theory is based on the real-valued preference functional V on $D[0,M]$, where $D[0,M]$ is the set of all distribution functions over the interval $[0,M]$. In our context, since the payment is either A or 0, we consider the set $D[0,A]$. Let $\Delta D[0,A] = \{\lambda(F^* - F)F, F^* \in D[0,A], \lambda \in \mathbf{R}\}$. We use the norm $||\cdot||$ on $\Delta D[0,A]$ such that $||\lambda(F^* - F)|| = |\lambda| \int |F^*(x) - F(x)|dx$. We assume that V is Fréchet differentiable on the space $D[0,M]$ with respect to the norm $||\cdot||$. This means that there exists a continuous liner functional $\psi(\cdot;F)$ defined on $\Delta D[0,A]$ such that

$$\lim_{||F^*-F||\to 0} \frac{|V(F^*) - V(F) - \psi(F^* - F;F)|}{||F^* - F||} = 0. \tag{1}$$

[9] In the theory of 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.

Under this assumption, there exists a absolutely continuous function $U(x;F)$ such that we can write

$$V(F^*) - V(F) = \int U(a;F)(dF^*(a) - dF(a)) + o(||F^* - F||). \tag{2}$$

The function $U(x;F)$ is called the local utility function. In the expected utility theory, it does not depend on F and is a usual utility function. This derivation shows that an agent with preference function V acts locally as if he was an expected utility maximizer. We obtain the desired result by assuming that the local utility function is increasing.

Assumption 2. *i) V is Fréchet differentiable on the space $D[0,A]$ with respect to the norm $||\cdot||$.*
ii) $U(a;F)$ is strictly increasing in x for all $F \in D[0,A]$

We note that $F(\cdot,\theta) \in D[0,A]$. The following theory shows that the BSR can elicit a property of the belief even under Machina's (1982) framework.

Theorem 3. *Suppose that Assumption 1i, 1ii and 1iv, and Assumption 2 hold. Assume that $l(\theta,X) < \overline{K}$ for any θ and X. Then,*

$$\arg\max_{\theta\in\Theta} V(F(\cdot,\theta)) = \arg\min_{\theta\in\Theta} E(l(\theta,X)|\mathcal{F}).$$

Proof. We first note that

$$E\left(\mathbf{1}_{\{l(\theta,X)<K\}}\middle|\,\mathcal{F}\right) = E\left(1 - \frac{1}{\overline{K}}l(\theta,X)|\mathcal{F}\right) = 1 - \frac{1}{\overline{K}}E\left(l(\theta,X)|\mathcal{F}\right). \tag{3}$$

Let $\theta^* \in \arg\min_{\theta\in\Theta} E(l(\theta,X)|\mathcal{F})$. Then $F(\cdot,\theta^*)$ stochastically dominates $F(\cdot,\theta)$ for $\theta \notin \arg\min_{\theta\in\Theta} E(l(\theta,X)|\mathcal{F})$. Theorem 1 in Machina (1982) shows that $V(F(\theta^*)) > V(F(\theta))$ for $\theta \notin \arg\min_{\theta\in\Theta} E(l(\theta,X)|\mathcal{F})$. (We note that Theorem 1 in Machina (1982) shows the weak inequality by assuming $U(x;F)$ is non-decreasing in x and asserts that assuming strict monotonicity ensures the strict inequality version of the theorem.) □

However, there are also decision theories under which monotonicity with respect to stochastic dominance is not satisfied. A notable example of this kind of theory is the prospect theory by Kahneman and Tversky (1979). Arguably, prospect theory provides the most popular model of non-expected utility in the literature. We show that the BSR can be used to elicit the belief under the prospect theory framework with some assumption. To explain the framework, we consider the following lottery: giving B with probability $\Pr(B)$ and giving A with probability $\Pr(A) = 1 - \Pr(B)$. Prospect theory surmises that an agent evaluates her utility from this lottery as

$$w(\Pr(A))u(A) + w(\Pr(B))u(B),$$

where $w(\cdot)$ is the probability weighting function with $w(0) = 0$ and $w(1) = 1$. We assume that this function is strictly increasing.

Assumption 3. *The probability weighting function is strictly increasing,* $w(\cdot) : [0,1] \to [0,1]$.

When w is the identity function, it is equivalent to expected utility. However, in general, it yields a different prediction about the decision of an agent from that under the expected utility framework because her evaluation of the probability is skewed by the probability weighting function, w. On the other hand, this framework satisfies "probabilistic sophistication" in the sense that the agent can understand and evaluate probability correctly. She distorts probability only when she evaluates the expected value of the utility from a lottery. Belief can be well-defined in this framework (it may not be well-defined in a more general model for decision under uncertainty) and it is sensible to ask a property of a random variable according to an agent's belief.

We need the following assumption on the utility function.

Assumption 4. $u(A) > 0$ *and* $u(B) \leq 0$.

This assumption implies that winning the lottery is considered as a positive prospect and an agent sees it as a gain. More importantly, it also assumes that losing the lottery is regarded as a negative prospect and she thinks that it is a loss. This assumption may be restrictive. However, the BSR may not be able to elicit a property of a subjective probability without this assumption. A practical implication of this assumption is that we should not give any reward, or should give a penalty when an agent loses the lottery. On the other hand, the BSR may not be used in a design in which she receives some reward which is smaller than the reward for winning when she loses the lottery.

The following theorem shows that the value of θ that minimizes the expected loss function $l(X, \theta)$ also maximizes the (non-expected) utility. It justifies the use of the binarized scoring rule under the prospect theory framework.

Theorem 4. *Suppose that Assumptions 1, 3 and 4 hold. Assume that $l(X, \theta) < \overline{K}$ for any X and θ, Then,*

$$\begin{aligned} & \arg\max_{\theta\in\Theta} \left(w(\Pr(l(X,\theta) < K|\mathcal{F}))u(A) + w(\Pr(l(X,\theta) \geq K|\mathcal{F}))u(0)\right) \\ = \;& \arg\min_{\theta\in\Theta} E(l(X,\theta)|\mathcal{F}). \end{aligned}$$

Proof. Since w is strictly increasing in its argument and $u(A) > 0$ and $u(0) \leq 0$, it follows that

$$\begin{aligned} & \arg\max_{\theta\in\Theta} \left(w(\Pr(l(X,\theta) < K|\mathcal{F}))u(A) + w(\Pr(l(X,\theta) \geq K|\mathcal{F}))u(0)\right) \\ = \;& \arg\max_{\theta\in\Theta} \left(w(\Pr(l(X,\theta) < K|\mathcal{F}))\right) \\ = \;& \arg\max_{\theta\in\Theta} \Pr(l(X,\theta) < K|\mathcal{F}). \end{aligned}$$

The rest of the proof follows by the same argument used in the proof of Theorem 1. □

The idea of the proof is following: Because the agent receives the reward that brings a positive utility only when the loss is less than K, maximizing her utility is equivalent to maximizing the weighted probability of the loss being less than K and minimizing the weighted probability of the opposite event. The probability weighting function is strictly increasing so that it is equivalent to maximizing the probability of receiving the reward. The maximization of the probability is, as shown in the proof of Theorem 1, equivalent to the minimization of the loss function.

In this theorem, we assume that the loss function is bounded. However, as in Theorem 2, we can relax the boundedness assumption with some additional assumptions that are specific to the loss function. For example, the validity of the binarized scoring rule for eliciting the mean can be shown exactly in the same way as in the proof of Theorem 2.

Offerman, Sonnemans, van de Kuilen and Wakker (2009, OSVW hereafter) also develop a method to elicit the probability of an event under the prospect theory framework. Their approach is based on the quadratic scoring rule and involves two steps. The first step is "calibration," in which the experimenters ask the agent the probability of many events with known probability using the quadratic scoring rule. In the second step, they ask the probability of an event of interest using the quadratic scoring rule and then correct the bias in the answer using the result of the calibration step. The binarized scoring rule has several advantages over the method of OSVW. First of all, our method can reveal the belief in one step; it does not require the calibration step. Since the calibration step may take time and it is not clear whether we can implement the calibration step in all circumstances, we believe that this advantage is important in practice. This advantage may also be important from a view of recent developments in decision theory. It may not be guaranteed that the utility function and the probability weighting function do not depend on the context. So it may be the case that the utility function in the calibration step is different from that in the second step, and if it is so, the method of OSVW cannot be applied. On the other hand, since our method is a one-step procedure, our method does not suffer from this problem. The second advantage is that the BSR does not need to make any assumption on the structural form of the utility function (or the probability weighting function) unlike as in OSVW. The third advantage is that the binarized scoring rule is much more widely applicable. It can be used not only to elicit the probability of an event but also to elicit many other properties of an random variable, such as the mean or the median. On the other hand, it is not clear how to extend the method of OSVW to directly elicit properties of a random variable other than probability.

Another approach is that of Andersen, Fountain, Harrison, and Rutström (2010) who consider estimating the subjective belief and the utility function jointly. They allow non-expected utility, in particular, they consider the rank dependent utility theory. Again the advantage of our approach is its simplicity. We do not need many data points to jointly estimate the utility function, rather we need to ask just one question using the binarized scoring rule to elicit a property, which we are interested in, of the random variable. However, an advantage of the joint estimation approach is that one can statistically analyze the elicited belief using that approach.

In addition to being very general and flexible, the binarized scoring rule is simple to explain and use. To illustrate how this scoring rule can be utilized in eliciting beliefs in a realistic setting, we ran laboratory experiments where we used both our scoring rule and the standard quadratic scoring rule to incentivize subjects. The experimental design and the results from the experimental sessions are discussed in the following section.

3 Experimental Illustration of the BSR

We applied the theoretical results in an experimental framework to analyze belief elicitation about the realized value of an unknown variable. Using the incentive scheme suggested in Theorems 1 and 2, we ran two types of experiments in which we elicit subjects' beliefs about certain aspects of a random variable. In the first type of the experiments, which we call the P-experiment, we elicit subjects' estimates of the probability that the ball drawn from an urn is of some specified color using both the quadratic and the binarized scoring rules. In the other two sessions, which we call 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 under the two scoring rules.

3.1 Experimental Design

The experiments presented in this paper can be divided into two groups—P-experiments and M-experiments. We elicited subjects' belief of an event happening in the P-experiments run in Hokkaido University, Japan and we elicited subjects' belief about the realized value of random variable in the M-experiments run in the Hong Kong University of Science and Technology. The experiments were programmed and conducted with the software *z-Tree* developed by Fischbacher (2007). All of the subjects were undergraduate students of the respective institutions and were recruited using a database of students willing to participate in economic experiments.

In the P-experiment, 153 subjects participated. The subjects were asked to predict the event that a ball randomly drawn from an urn with 100 balls of three different colors—red, blue, and black is of a certain color. To ensure that the subjects are aware of the concept of probability, they first reviewed the probability rules in this simple setting. Then, they participated in 10 practice periods. In each period, the color composition of the balls in the urn was different. They were asked to report a number to represent their prediction about an event concerning the color of the ball under both scoring rules—BSR and QSR. The subjects were not paid for these 10 periods. However, they were informed of the outcome of the draw and how much they would have earned for their predictions under each scoring rule after each period. Then, they participated in two paid periods. In one period, they were paid using the BSR and in the other period they were paid using the QSR. The color composition of the urn, the event which the subject had to predict in each round and which incentive scheme was used first was randomly decided by the computer program for each subject. Subjects were clearly informed of the color composition of the 100 balls in the urn at the beginning of each period. Moreover, they were informed of the outcome of the draw and their income from both rounds 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 in the period in which she was paid using the QSR depended on her exact utility function. Before participating in ten practice periods and two paid periods, the subjects also participated in a five-period round designed to glean information about their risk-preference. In each of these periods, they were presented with a lottery and they had to report their certainty equivalent for the 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 Becker-DeGroot-Marschak (1964) mechanism. At the end of the session, they were paid in cash for the two paid periods and one 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 are provided with copies of the sheet that shows that the relationship between their report and the payment.[10]

Each subject started with an initial endowment of JPY 1000. Subjects were asked to report their prediction about the drawn ball being of a certain

[10] These can be supplied by the authors upon request.

color or not of a certain color in terms of a number. They could enter any integer between 0 and 100 (inclusive) to denote their prediction of the relevant event happening. Suppose in a period, a subject was asked to supply her prediction that the color of the drawn ball was red and she entered a number P. 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 600 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 200 would be taken away from her endowment. Under the QSR incentive scheme, the subject received JPY $600 - 800sqe$ where sqe stands for the squared error. If a subject reported the true objective probability (in percentage terms) as her prediction (the optimal choice under risk-neutrality), her expected earning would be the same under both incentive schemes. Subjects spent around an hour in the laboratory on average and the average payment to each subject was around JPY 1745, which equaled slightly below USD 21, at the exchange rate during the time of the experiment.

In the M-experiment, 62 subjects participated in two sessions, each of which consisted of 40 periods. Subjects earned points in each period and the total points earned were translated into Hong Kong dollar (HKD) according to an exchange rate specified at the beginning of the experiment. The average payment to subjects was slightly above HKD 164 for a time period of around an hour and a half, considerably higher than wages from outside work options available to most subjects.[11] They were paid in cash at the end of the session. The sessions were conducted in English and the experimental rules were described using PowerPoint presentations. Subjects were provided with copies of the presentation slides and a note on definitions of relevant

[11] USD 1 $\approx$ HKD 7.80.

statistical concepts.[12]

To put signals about the realization of a random variable in a familiar context, subjects participated in experimental games involving assets that provide uncertain returns. A subject was endowed with a stock of a new (fictitious) company at the beginning of each period. The company in a period was independent of companies in other periods. A subject's earnings from different periods were, thus, independent. In period $t \in \{1, 2, \ldots, 40\}$, a subject received estimates or forecasts of the earning per share (EPS) from the stock t from 10 separate analysts. All 31 subjects in the session received the same set of 10 forecasts. These forecasts were independently drawn according to a process that was explained thoroughly at the beginning of the sessions. Suppose that the true EPS of this stock was T. Analyst $i \in \{1, 2, \ldots, 10\}$ reported a forecast $F_i = T + \epsilon_i$ where ϵ_i was drawn from a normal distribution with mean zero and variance σ_F^2 and $\epsilon_1, \ldots, \epsilon_{10}$ were mutually independent. Hence, all forecasts were unbiased. Subjects were informed of the value of the parameter σ_F^2 and the forecast and EPS generating processes. In each period, a new stock (with a different EPS) was presented.

Armed with the forecasts, each player made predictions about the realized earning per share from that period's stock. A subject was asked to enter her estimate of the true EPS. Suppose the predicted realization of the EPS that a subject entered was M and the realized value was T. Then the subject's squared prediction error equals $(T - M)^2$. In these sessions, the EPS of a stock in each period was independently drawn from a normal distribution with mean 60 and variance 400. Moreover, the variance of the distribution that generated the error term (σ_F^2) was 8. Under the BSR, if the squared error was below some number K, she won a fixed prize of 80 points. In each period, a new error bound K was generated from a uniform distribution on $\left[0, \overline{K}\right]$. The subject was informed of the realized values of K and the EPS (T) for a

[12] These can be supplied by the authors upon request.

given period only at the end of that period. For these sessions, we chose $\overline{K}$ to equal 6 so that Theorem 2 may be applied. By calculating the distribution of the posterior mean, we estimate that the theoretical probability of the error being above 6 is about 0.6% when a subject takes the optimal action.[13] Theorem 2 implies that the optimal strategy for a subject in these sessions under BSR was to choose the prediction that minimizes the expected mean squared error given her forecasts or signals irrespective of her risk-attitude. That is, her optimal prediction is the value v that minimizes $E\left[(T-v)^2\right]$ given the 10 forecasts she received irrespective of her preferences as long as her utility from a monetary prize is positive. As all the forecasts are independent, a subject's optimal strategy is to report the posterior mean of the true EPS given the 10 forecasts she received irrespective of her risk-preference. Given that the true EPS is drawn from $N(60, 400)$, the posterior is distributed according to $N(500\bar{X}/501 + 60/501, 400/501)$, where $\bar{X}$ is the mean of the 10 forecasts. Therefore, one can easily approximate a subject's optimal policy under the BSR by the mean of the 10 forecasts she receives.

Under the QSR, the subject received a payment of $90 - 25\,(T-M)^2$. We chose the parameters in QSR (i.e., 90 and 25) such that the mean and variance of earnings when subjects followed the optimal rule are the same for both scoring rules. In the periods when this value was negative for a subject, she received negative points for that period. Interestingly, as Bhattacharya and Pfleiderer (1985) showed, a weakly risk-averse subject's optimal strategy is to report the mean of the 10 forecasts she received even under QSR as the realization of the stock return is derived from a symmetric distribution when the utility function satisfies some regularity conditions. Thus, in the M-experiment, virtually the same behavior (reporting the mean) is predicted under both BSR and QSR. Lastly, the exchange rate was 16 points = 1 HKD in the M-experiment.

[13] Among the 2436 observations, only 60 observations had squared errors larger than 6.

In the first of the two sessions, subjects were incentivized using the BSR in the first twenty periods and were incentivized using the QSR in the remaining twenty periods. In the second session, QSR was used in periods 1 to 20 and BSR was used in periods 21 to 40. As mentioned earlier, the realized value and generated forecasts were different in the forty periods. In the context of the experimental setting, the stocks were different in the forty periods. However, we used the same set of forty stocks and corresponding forecasts in the two sessions. Comparing the performance of a given subject in the first 20 and the last 20 periods, we can compare the two scoring rules for that subject, albeit using different stocks. On the other hand, comparing results for a given period between the two sessions, we thus can compare the two scoring rules for the same stock, albeit using different subjects.

3.2 Results

In this section, we describe and analyze our experimental results. Using the two sets of experiments, we compare the performances of the two scoring rules—BSR and QSR. As described in Section 3.1, we ran two distinct types of experiments, the P-experiment and the M-experiment. The objective in the P-experiment is to elicit the belief about the probability of the event that the ball drawn from an urn is of some specified color. For the M-experiment, it is to elicit the belief about the mean of the unknown variable. Excluding some missing and abnormal observations, the numbers of observations used in our data analysis are 276 for the P-experiment and 2436 for the M-experiment.[14] We regard our data set as panel data and assume that the observations are independent across subjects but are potentially dependent over periods.

[14] We examine whether the likelihood of producing abnormal observations are associated with the incentive scheme used and we find no statistical evidence of this association. We also try several different criteria for abnormal observations and find that the results presented below hold qualitatively under any criterion.

To measure subject performance to compare the binarized and quadratic scoring rules, we compute the negative of the square of the difference between the reported number and the action that minimizes the expected loss. We denote this measure by NSD. We call the action that minimizes the expected loss the *optimal* action. It is the optimal action under the BSR scheme, independent of the subject's risk-preference, which is the same as her optimal action under QSR if she is risk-neutral. For the P-experiment, the optimal action is to report the objective probability of the specified event happening and, for the M-experiment, the optimal action is approximated by reporting the mean of the ten forecasts. Descriptive statistics of NSD are available in Table 1.[15]

3.2.1 P-Experiment

We examine the results from the P-experiment. Figures 1(a) and 1(b) plot the reported prediction against the true probability when the incentive scheme is the BSR and the QSR, respectively. We see a tendency that subjects reported numbers close to 50 when the QSR is used. We run regressions of NDS on BSR, which is a binary variable and takes one when the BSR is used and zero otherwise, other variables to examine the effect of incentive scheme on reported predictions. Table 2 summarizes the results of these regressions. Standard errors are heteroskedasticity and autocorrelation robust standard errors, which allows arbitrary heteroskedasticity and correlation within an individual but assumes independence across subjects (Arellano, 1987). Column 1 indicates that the reported predictions from the BSR are on average $8(=\sqrt{64})$ closer to the true probabilities in percentage than those from the QSR. In column 2, we control for the order in which the two incentive schemes were used in the paid belief elicitation periods. However, the result does not change substantially. The results are also robust to controlling for personal

[15] We find similar results when we use other measures of performance such as calculating the (hypothetical) reward under the same scoring rule, either BSR or QSR, for all periods.

characteristics such as gender, major, or class level of the subjects.

Another indication of the good property of the BSR is illustrated by the regressions presented in Table 3. We run regressions of the reported prediction on the true probability. When the reported prediction equals to the true probability, the intercept in this regression should be zero and the coefficient on the true probability is 1. When subjects submitted numbers close to 50 (as risk averse subjects do under the QSR according to the theory), the coefficient on the true probability is less than 1. Column 1 is from this regression with the whole sample. The result shows that subjects did not submit the true probability always and had submitted numbers close to 50. In column 2, we add BSR and the interaction term, which is the same as dividing the sample according to the value of BSR. We see that the subjects provided numbers close to the true probability under the BSR, while they provided numbers close to 50 under the QSR. We examine whether this difference between the BSR and the QSR comes from the risk averseness of the subjects. Columns 3 and 4 show the results from the regressions with a risk aversion measure and its interaction using the observations from the BSR and those of the QSR, respectively. The risk aversion parameter we use here is the number of period in which a subject submitted a certainty-equivalent below or equal to the expected value of the lottery in the risk-preference elicitation round. The results show that reported probabilities under the BSR are not affected by the risk aversion parameter while those under the QSR are. This finding is consistent with the theory that the optimal action under the BSR is independent of risk attitude while a risk-averse subject should submit a number close to 50 under the QSR.

3.2.2 M-Experiment

To compare the performances of subjects under the BSR and the QSR, we regress NSD on the binary variable BSR. Moreover, we try specifications that add the variable $EXPERIENCE$, which is the number of periods that the subject had experienced up to that period under that particular scoring rule, to see whether the behavior of subjects changes over time. It equals the period number in periods 1 to 20 and the period number minus 20 in periods 21 to 40. We also add the interaction between BSR and $EXPERIENCE$ to allow for different effects of experience between the two scoring rules.[16]

Table 4 summarizes the estimation results. Standard errors are heteroskedasticity and autocorrelation robust standard errors. The table shows that the differences between the performances under the two scoring rules are not statistically significant, nor economically important in the M-experiment. This result holds regardless of the choice of performance measure. Figures 2(a) and 2(b) are the scatter plots of reported number versus the optimal action (the optimal action is the mean of the 10 forecasts, $\bar{X}$, in this case) under the BSR and the QSR, respectively. There is no notable difference between the two plots. The observation from the figures confirms our statistical finding that the behavior under the QSR is not substantially different from that under the BSR. We then examine the effect of experience. We do not find any statistically significant effect of experience on performance. Although BSR is incentive compatible only at the limit as $\overline{K}$ approaches infinity. The fact that BSR performs as well as the QSR in the lab even with a $\overline{K}$ of 6, shows that BSR is quite effective in practice.[17]

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

[17] We also examine the strategies used by the subjects in the experiments. In particular, we regress $PREDICTION$ (the entered predicted value of the subjects for the true EPS of a given stock) on the mean, maximum, minimum and the standard deviations of the 10 forecasts the subjects received. Recall that the optimal strategy under risk-neutrality is very close to the mean of the forecasts. Empirically, we find that the weight on the mean of forecasts is statistically indifferent from 1. There is no statistically significant evidence that the maximum, the minimum, or standard deviation of analyst forecasts affect subjects' action. We do not find any evidence that subjects behaved differently under the two scoring rules either. These results are available from the authors upon request.

3.2.3 Summary

We examine the results of the experiments to investigate the difference in subjects' behavior under the two scoring rules. For the P-experiment, where risk-aversion theoretically does not lead to truthful revelation of the objective probability under the QSR, we find that subjects' behavior is closer to the optimal action for risk-neutral agents under the BSR. For the M-experiment, we do not find a statistically significant difference and the estimated differences are all small in terms of magnitude. This result is consistent with the theoretical prediction of Bhattacharya and Pfleiderer (1985). Our results are in stark contrast to the results of Selten, Sadrieh and Abbink (1999) who found that a randomized scoring rule does not work well compared to deterministic scoring rules. Thus, we provide a simple scoring rule that is theoretically more general and superior in eliciting beliefs and show that this rule performs better than the widely-used quadratic scoring rule in standard experiments of belief elicitation. This paves the road for much wider use of this scoring rule in more complicated situations.

4 Conclusions

This paper introduces a general mechanism to create incentive compatible scoring rules for eliciting subjective belief of economic agent without making any strong assumption on the agent's risk preference. We also show that the new scoring rule works even under non-expected theory frameworks. We observe that simple scoring rules using this mechanism perform quite well 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. As an example, a natural application can be situations where subjects receive many signals, some of which are independent while others are correlated, about the realized value of a random variable. Such an experiment can also be used to test how well people can differentiate relative importance of independent and correlated signals and is on the agenda for future research.

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

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

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

[6] Blanco, Mariana, Dirk Engelmann, Alexander K. Koch and Hans-Theo Normann (2010): “Belief Elicitation in Experiments: Is There a Hedging Problem?,” forthcoming in *Experimental Economics.*

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

[8] Chambers, Christopher P., (2008): “Proper Scoring Rules for General Decision Models,” *Games and Economic Behavior*, 63, 32-40.

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

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

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

[12] Friedman, Daniel (1983): "Effective Scoring Rules for Probabilistic Forecasts," *Management Science*, 29(4), 447-454.

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

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

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

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

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

[18] 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.

[19] 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.

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

[21] Kahneman, Daniel and Amos Tversky (1979): "An Analysis of Decision under Risk," *Econometrica*, 47(2), 263-291.

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

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

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

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

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

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

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

	P-experiment	M-experiment
Average of NSD	-138.203	-0.218
(standard deviation)	279.145	0.584
# of observations	276	2436
Average of NSD under the BSR	-105.891	-0.207
(standard deviation)	175.994	0.523
# of observations	137	1216
Average of NSD under the QSR	-170.050	-0.230
(standard deviation)	350.359	0.639
# of 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 t-test statistics 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 the experiment results

Dependent variable	NSD	NSD
Constant	-170.050***	-78.054*
	(29.718)	(42.225)
BSR	64.160**	61.596**
	(29.689)	(29.204)
Order Effect	No	Yes
Adjusted R^2	0.010	0.017
# of observations	276	276

Note: "***", "**" and "*" indicate the 1%, 5% and 10% significance, 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); BSR is binary and = 1 if the BSR is used.

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

Dependent variable	Reported Prediction	Reported Prediction	Reported Prediction	Reported Prediction
Constant	5.683***	8.334***	2.076	-1.834
	(1.518)	(2.086)	(5.572)	(6.014)
Trueprob	0.876***	0.840***	0.983***	1.054***
	(0.027)	(0.036)	(0.110)	(0.094)
BSR		-5.851***		
		(2.194)		
BSR×Trueprob		0.080**		
		(0.039)		
RiskAversion			0.090	2.642*
			(1.395)	(1.567)
RiskAversion×Trueprob			-0.016	-0.056**
			(0.026)	(0.025)
Adjusted R^2	0.857	0.859	0.885	0.836
F-test: $\beta_{BSR} = 0$		3.587		
and $\beta_{BSR \times Trueprob} = 0$		(0.290)		
F-test: $\beta_{RiskAversion} = 0$ and			0.149	2.685
$\beta_{RiskAversion \times Trueprob} = 0$			(0.862)	(0.072)
F-test: $\beta_0 = 0$	10.701	9.863		
and $\beta_{Trueprob} = 1$	(0.000)	(0.000)		
F-test: $\beta_0 + \beta_{BSR} = 0$ and		4.513		
$\beta_{Trueprob} + \beta_{BSR \times Trueprob} = 1$		(0.012)		
BSR or QSR	Both	Both	BSR	QSR
# of observations	276	276	137	139

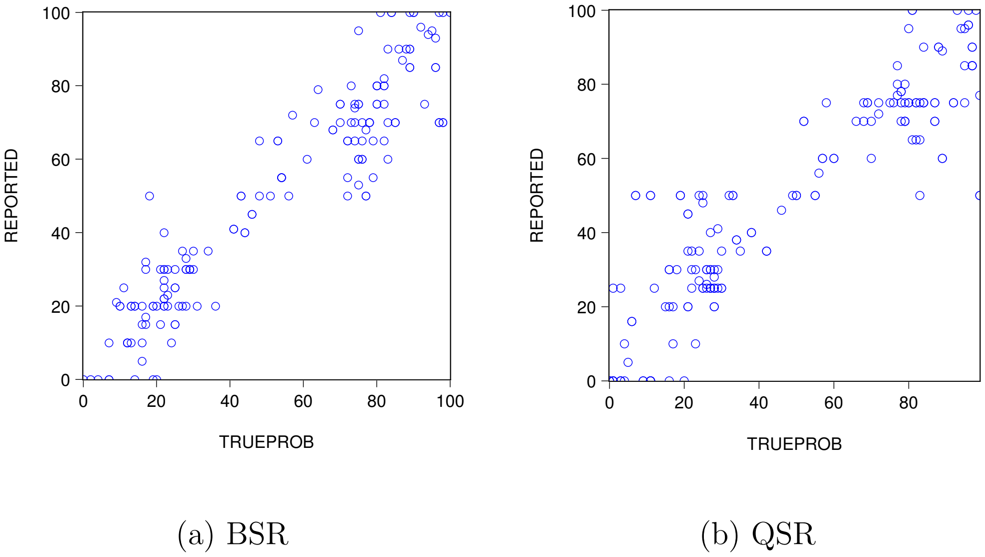
Note: "***", "**" and "*" indicate the 1%, 5% and 10% significance, respectively. Heteroskedasticity and autocorrelation robust standard errors are in parentheses. Trueprob is the objective probability in percentage terms; BSR is binary and = 1 if the BSR is used; RiskAversion is the measure of risk attitude defined in the main text. In parentheses under the values of the test statistics are p-values.

Table 3: Subject performance under the BSR and the QSR and its relationship with risk attitude, 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 × EXPERIENCE		0.000
		(0.005)
Adjusted R^2	0.000	0.003
F-test: $\beta_{EXPERIENCE} = 0$		2.403
and $\beta_{BSR \times EXPERIENCE} = 0$		(0.091)
F-test: $\beta_{BSR} = 0$		0.572
and $\beta_{BSR \times EXPERIENCE} = 0$		(0.565)
F-test: $\beta_{EXPERIENCE}$		2.938
$+\beta_{BSR \times EXPERIENCE} = 0$		(0.087)
sample size	2436	2436

Note: "***" indicates the 1% significance. Heteroskedasticity and autocorrelation robust standard errors are in parentheses. In parentheses under the values of the test statistics are p-values.

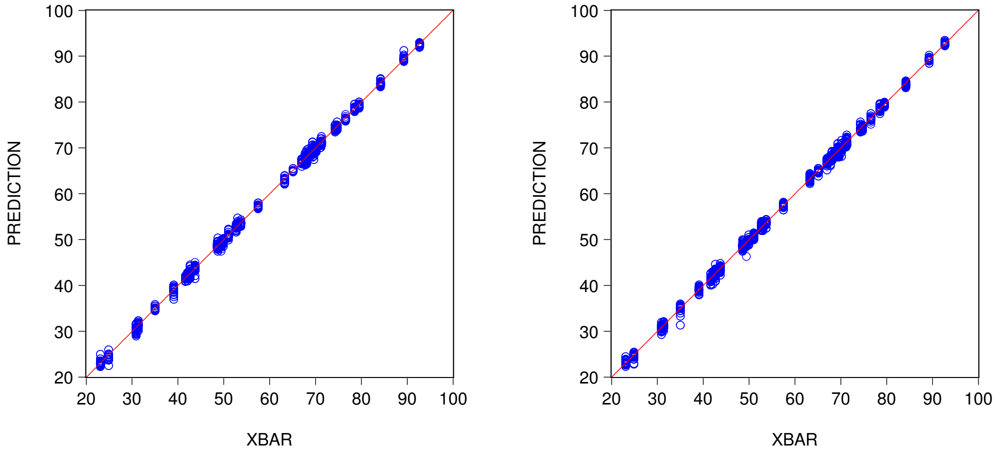
Table 4: Subject performance under the BSR and the QSR, M-experiment



(a) BSR (b) QSR

Note: These figures are the scatter plots of the reported prediction against the true probability in percentage under the binarized scoring rule (BSR) (left) and under the quadratic scoring rule (QSR) (right). Reporting the true probability is the optimal action under the BSR and, under risk neutrality, it is the optimal action under the QSR.

Figure 1: Reported vs true probability under in the P-experiment



(a) BSR

(b) QSR

Note: These figures are the scatter plots of the prediction against the value of the average of forecasts under the binarized scoring rule (BSR) (left) and under the quadratic scoring rule (QSR) (right). The line is the OLS regression line. Reporting the average of forecasts is the optimal action both under the BSR and the QSR.

Figure 2: Prediction vs $\bar{X}$ in the M-experiment