A Penny for Your Thoughts: A Survey of Methods for Eliciting Beliefs

Schlag, K.; Tremewan, J.; van der Weele, J.

Publication date
2013
Document Version
Final published version

A Penny for Your Thoughts: A Survey of Methods for Eliciting Beliefs.*

Karl Schlag† James Tremewan‡ Joël van der Weele§

November 12, 2013

Abstract

Incentivized methods for eliciting subjective probabilities in economic experiments present the subject with risky choices or bets that encourage truthful reporting. We discuss the most prominent elicitation methods and their underlying assumptions, provide theoretical comparisons, and propose some extensions to the standard framework. In addition, we survey the empirical literature on the performance of these elicitation methods in actual experiments, considering also practical issues of implementation such as order effects, hedging, and different ways of presenting probabilities and payment schemes to experimental subjects. We end with some thoughts on the merits of using incentives for belief elicitation and some guidelines for implementation.

JEL-codes: C83, C91, D83.

Keywords: belief elicitation, subjective beliefs, scoring rules, experimental design.

*We would like to thank Peter Wakker, Theo Offerman and Glenn Harrison for useful comments.

†University Vienna, Vienna. E-mail: karl.schlag@univie.ac.at

‡University Vienna, Vienna. E-mail: james.tremewan@univie.ac.at

§Corresponding author. University of Amsterdam. Email: vdweele@uva.nl. Tel. +31 (0)20 5254213. Address: CREED, Department of Economics, University of Amsterdam, Roeterstraat 11, 1018WB Amsterdam, the Netherlands.

1 Introduction

The beliefs of experimental subjects are often of great interest to experimentalists, as choice data alone are often not sufficient to distinguish between different theories (Manski, 2002, 2004). For example, first movers in ultimatum bargaining experiments may make high offers because they are altruistic, or because they believe the other party will reject low offers. Eliciting beliefs can help disentangle these hypotheses. Belief measurement is also necessary when beliefs are an object of study in themselves, as in experiments about expectation formation and updating.

Experimental economists have developed a set of practices of belief elicitation that differ from those of other disciplines. One characteristic that is typical (although not exclusive) to the economic approach is that beliefs are most commonly elicited as probabilistic statements. The underlying assumption is that beliefs take the form of subjective probabilities, or if they don't, can be usefully expressed in this form. As Manski (2004) points out, probabilistic statements have advantages over other formats. Since they have a well-defined scale, probabilistic statements allow for comparisons with objective frequencies and for an assessment of interpersonal heterogeneity. In addition, the logic underlying probabilistic reasoning makes it possible to evaluate the consistency of belief statements.

A second characteristic of economists' method's of belief elicitation is the use of rewards for reporting beliefs that turn out to be correct. This practice stems from the old idea that subjective probabilities are best measured by offering people bets with varying odds (Ramsey, 1926). The idea is that well designed bets give the subject an incentive to take the task seriously and report truthfully. For example, (risk averse) people will accept low odds on the occurrence of some event only when they are relatively sure the event will indeed occur. The design of such bets has been an active research field in economics for the past few decades.

In this article, we survey betting techniques that have been developed to elicit subjective probabilities from experimental subjects. The aim is to give experimentalists an overview of the available methods, their underlying assumptions and their empirical performance. In Section 2 we outline more formally the elicitation environment and discuss a class of incentive schemes called scoring rules. Scoring rules reward subjects on the basis of the submitted report, and the actual realization of the random variable. In Section 2.2 and 2.3 we discuss so-called 'proper scoring rules' that have been designed to elicit probabilities, means, modes and various quantiles truthfully, under the assumption that the subject is risk neutral. In Sections 2.4, 2.5 and 2.6 we discuss mechanisms that abandon this restrictive assumption. In Section 2.7 we discuss promising extensions to the standard framework.

We evaluate belief elicitation mechanisms empirically in Section 3, by looking at the 'quality' of elicited beliefs. Such evaluation is not straightforward, since by the nature of the exercise we do not know the right benchmark, i.e. the true belief of the subject. We discuss several alternative benchmarks that may be used to assess the effectiveness of elicitation mechanisms. In Section 4 we discuss practical issues in implementation such as the complexity and presentation

of incentive schemes. Finally, in Section 5 we tie our findings together and present some thoughts about the appropriate use of incentives in belief elicitation, and directions for future research on this topic.

Our survey is complemented by other reviews of the broader aspects of belief elicitation. Manski (2004) focuses on questionnaires in consumer and household surveys. In this context, incentivized elicitation is typically not possible, since beliefs cannot be immediately verified. Garthwaite *et al.* (2005) and Jenkinson (2005) present a survey of mostly unincentivized elicitation techniques with a wide range of applications. Delavande *et al.* (2011) consider belief elicitation in studies with subjects in developing countries who may have low levels of numerical literacy. Gneiting and Raftery (2007) present a technical review of proper scoring rules, including a detailed discussion of decision theoretic foundations and statistical properties. Winkler (1996) focuses on a limited set of proper scoring rules and discusses scoring rules as a tool for ex-post evaluation of forecasts, an issue we ignore in this survey.

2 Scoring rules

In this section, we discuss mechanisms used to incentivize truthful belief elicitation. We consider an 'experimenter' who wishes to learn something about the beliefs that a 'subject' holds about some future event. Common examples include learning about the probability that an event occurs, or about the expectation of some value of that will be realized in the future. Below we formalize this setting and focus on a class of mechanisms known as scoring rules: payment functions that depends on the report of the subject and the realization of the event.

2.1 Preliminaries

Consider an experimenter who is interested in a quantity that is related to the beliefs of the subject about the distribution of a random variable X. For instance, the experimenter is interested in the probability that $X = 1$, or in the expected value of X. Let $\mathcal{X}$ denote the set of possible realizations of X. To simplify exposition we assume that $\mathcal{X}$ is finite, which is typically satisfied in experiments. For instance, if x is an integer value on a scale that ranges from 1 to 10, then $\mathcal{X} = \{1, ..., 10\}$. Let P_X be the probability distribution that describes the beliefs of the subject, so for each value x belonging to $\mathcal{X}$ it describes the probability that the subject believes $X = x$. The experimenter does not know P_X and wishes to learn about some parameter (or property) θ of this belief distribution. Common examples for θ are the probability of an event and the expected value or median of a future realization of X. Note however that θ can also be a more sophisticated object, such as a 95% confidence interval for x such that the probabilities of x falling below and above the interval are equal. Let $\mathcal{P}_X$ be the set of possible belief distributions of the subject and Θ the set of possible values of θ. Formally, θ is a mapping from the set of

distributions into Θ.[1]

It is natural to ask the subject to report θ directly. In fact, there is no loss of generality to limit our presentation to payment schemes where the subject is asked to report θ. This follows from the same arguments that lead to the revelation principle in mechanism design. In what follows, we denote the subject's report by r and her true subjective belief by p.

To simplify elicitation, one typically only considers payment based on a single realization of X. This is the case we consider here. Settings where it is useful or even necessary to condition on multiple realizations are discussed in Section 2.7. So the payment scheme S, which is also called a *scoring rule*, is a mapping from $\Theta \times \mathcal{X}$ into $\mathbb{R}$, where $S(r, x)$ is the amount of money paid to an expert when outcome x is realized after the expert has reported r. We assume that the realization X is independent of the report r of the subject.

To predict the effect of incentives on reporting behavior, we have to specify the decision making process of the subject. For now, we stay within the canonical model of decision making, and assume that the subject is an expected utility maximizer. Alternative models of decision making are discussed later. Suppose the subject has some utility function u, and given the payment function S reports an element of

$$\arg\max_{r \in \Theta} Eu\left(S\left(r, X\right)\right)$$

where

$$Eu\left(S\left(r, X\right)\right) = \sum_{x \in \mathcal{X}} u\left(S\left(r, x\right)\right) P\left(X = x\right)$$

is the expected utility of reporting r.

Now consider the experimenter, who wishes to design the incentives such that they induce the subject to tell the truth. A first issue is that the experimenter may not know the utility function of the subject. Let U be the set of possible objective functions of the subject as assessed by the experimenter. If U contains a single element then it is as if the experimenter knows the objective of the subject. For instance, if $U = \{Id\}$, where $Id(x) = x$ for all x, then we are considering the case where the experimenter believes that the subject is risk neutral.

We call a scoring rule 'truth-telling' if it induces the subject to tell the truth, regardless of which utility function $u \in U$ the subject is basing her choices on.[2] More formally, S is called a *truth-telling rule for θ for subjects that have utility belonging to U* if

$$\{\theta(X)\} = \arg\max_{r \in \Theta} Eu\left(S\left(r, X\right)\right) \text{ for all } u \in U \text{ and all } P_X \in \mathcal{P}_X.$$

[1] We assume that θ is uniquely determined given X, so $\theta = \theta(X)$. Definitions become a bit more involved when the parameter of interest is not always uniquely defined, see Section 2.7.

[2] Truth-telling applies the concept of incentive compatibility to belief elicitation. It is noteworthy that the earliest incentive compatible scoring rules were proposed about two decades before the first work on mechanism design was published.

We say that θ *can be elicited for the subjects with utility belonging to* U if there is a scoring rule S that is truth-telling rule for U.

2.2 Proper scoring rules: Elicitation when subjects are risk neutral

For the special case in which the subjected is assessed to be risk neutral (i.e. $U = \{Id\}$) it turns out the most common parameters can be elicited. Starting at least with Winkler and Murphy (1968) the literature refers to such rules as *proper scoring rules* (PSRs). Below, we characterize the general mathematical representations of such rules and consider specific examples. Unless stated otherwise, the methods below will not be truth-telling if the subject is not risk neutral.

Lambert *et al.* (2008) give a general characterization of what can be elicited using a single realization when the subject is risk neutral. Assume that the experimenter wishes to elicit θ where $\theta = \theta(X)$ is continuous and not constant on any open neighborhood. Then θ can be elicited if and only if θ^{-1} is convex and maximal within the set of possible reports. This immediately implies that the mean, any moment and any quantile is elicitable, but that the variance cannot be elicited. The variance can be elicited when two realizations are available (see Section 2.7).

2.2.1 Eliciting probabilities of events

We outline two general representations of proper scoring rules for the elicitation of probabilities of events, due to Savage (1971) and Schervish (1989).[3] Below, we use the representation of Schervish to justify the quadratic rule, and the representation of Savage to justify the logarithmic rule.

Savage (1971, Section 6.1) gives the following general characterization. Given report r, assume that the payment to the subject equals $Y(r)$ if the event occurs and $Z(r)$ if it does not occur. So $ES(r,p) = Y(r)\,p + Z(r)\,(1-p)$. Then S elicits the probability of an event if and only if $J(p) = ES(p,p)$ is strictly convex in p and the graph of $r \to ES(r,p)$ is tangent to the graph of $r \to ES(r,r)$ for all p. In particular, $J(p)$ must be differentiable almost everywhere with $J'(p) = Y(p) - Z(p)$ whenever it is differentiable.

Schervish (1989, Theorem A9) gives the following characterization. Let S be a scoring rule for eliciting the probability of an event where $S \geq 0$ and S is continuous at the boundaries $\{r,x\} \in \{0,1\}^2$. Then S is strictly proper if and only if there exists a nonnegative measure ν with at most countably many point masses that assigns positive measure to every open interval

[3]Early work can be found in McCarthy (1956, Theorem 1) and Shuford *et al.* (1966, Theorem 1).

such that[4]

$$\begin{aligned} S(r,1) &= S(1,1) - \int (1-c)\,\mathbf{1}_{\{r\leq c\}}\nu(dc) \\ S(r,0) &= S(0,0) - \int c\mathbf{1}_{\{r>c\}}\nu(dc). \end{aligned}$$

Note that any convex combination of two proper scoring rules and a positive affine transformation of a proper scoring rule is also a proper scoring rule (see Gneiting and Raftery, 2007, Section 6). We now consider a number of applications of this general framework.

The quadratic scoring rule. By far the most used and well-known rule to elicit probabilities is the quadratic scoring rule (QSR). This rule is based on the Brier score (Brier, 1950), which is the sum of the squared errors of the reported probabilities. Let there be n possible outcomes or events, indexed by $i = 1, 2, ..., n$, with associated reports r_i. The Brier score is given by

$$S(\mathbf{r}, x) = -\sum_{i=1}^{n} (\mathbf{I}_i - r_i)^2, \tag{1}$$

where $\mathbf{I}_i$ is 1 if $x \in E_i$ and 0 otherwise. It is usual to take half the value of this original Brier score, so that the resulting score lies between 0 and 1. Thus, when there are two outcomes, the Brier score is $-(1-r)^2$ if the event occurs and $-r^2$ when it does not.

Since any proper scoring rule remains proper under an affine transformation, the Brier score can generalized. If event j occurs, this general version of the QSR is given by

$$S(\mathbf{r}, x) = a + b\left(2r_j - \sum_{i=1}^{n} r_i^2\right), \tag{2}$$

where a and b can be set by the experimenter. A common choice is to set $a = b$, so payoffs fall in the range $[0, 2a]$. The QSR punishes the subject according to the square of the distance between the specified probability and the actual outcome. This rule is strictly proper for $\theta = \Pr(X \in E_i)$.

When there is a single event of interest, one can implement the QSR using the following mechanism. The subject faces a price q that is continuously increasing, starting at 0. At each price q subject is buying 2 units of probability of obtaining prize $y = 1$ if the event occurs. The subject is asked to report the price r at which she wants to exit the market and stop buying. To see that this mechanism implements the QSR, note that the if subject exits at r the probability of getting prize is $\int_0^r 2dq = 2r$, while her payment is $\int_0^r 2qdq = r^2$. In other words, $S(r,1) = 2r - r^2$ and $S(r,0) = -r^2$, which is equal to the QSR.

[4] $\mathbf{1}_{\{r\leq c\}} = 1$ if $r \leq c$ and $= 0$ if $r > c$. $\mathbf{1}_{\{r>c\}}$ is defined similarly.

The spherical scoring rule. The spherical scoring rule, due to Roby (1964), is a strictly proper scoring rule given by

$$S(\mathbf{r}, x) = -\frac{r_j}{|\mathbf{r}|} = -\frac{r_j}{\sqrt{\sum_{i=1}^{n} r_i^2}} \tag{3}$$

Thus, this rule pays according to the relative probability specified for the event that occurred. Selten (1998) provides a proof that this rule is strictly proper, and also notes that the sensitivity of the spherical scoring rule depends on the factor $\frac{1}{|r|}$, which is maximized when $r_1 = r_2 = ... = r_n$. Thus, the spherical rule provides the strongest incentives to tell the truth when events are thought to be equally likely.

The logarithmic scoring rule. The logarithmic scoring rule (Good, 1952; Toda, 1963) is given by

$$S(\mathbf{r}, x) = -\ln r_j. \tag{4}$$

This rule is strictly proper and has the appealing property that it depends only on the probability assigned to the correct answer, not on those assigned to the other, incorrect answers (i.e. $S(r, i) = f(r_i)$). Note that the logarithmic scoring rule itself is unbounded: when an event occurs that the subject predicted to be impossible ($r_i = 0$) the score is $-\infty$. Thus, the rule needs to be truncated for experimental practice (Shuford *et al.*, 1966), but will no longer be strictly proper after such a truncation.[5]

Certainty equivalents. De Finetti (1970) and Savage (1971) note that a probability can be viewed as a prize, or a marginal rate of substitution from probabilistic to sure payoffs. This opens another avenue to probability elicitation that is based on the a reservation-price elicitation mechanism of Becker *et al.* (BDM, 1964). In the so-called "promissory note" method, described in De Finetti (1974, but see also Ramsey (1926)), the experimenter asks the subject to report the lowest price r she would be willing to pay to acquire a prospect $y_E g$ (i.e. a lottery that pays y if event E occurs and g otherwise). Typically $g = 0$, so that the lottery simply pays y if the event occurs.

To determine the payment, a price z is randomly chosen according to the realization of a random variable Z that has distribution P_Z with support $[0, \infty)$. The subject receives the lottery if and only if the price $z < r$. If she is risk neutral, it is optimal for the subject to state her true certainty equivalent CE of the lottery: if she reported $r < CE$ she would be worse off when z falls in [r,CE). Similarly, reporting $r >$CE backfires when z is in (CE, r]. For a risk neutral

[5]Selten (1998) criticizes the logarithmic rule and shows that it is at the same time very sensitive to small mistakes for small probabilities and insensitive to the distance from the truth for predictions $r_i = 0$.

subject, the elicited certainty equivalent can be used to calculate the probability as $p = \frac{r-g}{y-g}$.[6]

2.2.2 Eliciting means

A general framework for proper scoring rules for the mean can be found in Savage (1971, Sections 6.2, 6.3), and is similar to that for eliciting events. An application of this framework is the QSR, which can be used to elicit the expected value or mean. The scoring rule in this case is given by

$$S^{QSR}(r,x) = a - b(r-x)^2. \tag{5}$$

where $a, b > 0$.

2.2.3 Eliciting the mode

To elicit the mode of a discrete distribution, it suffices to reward the subject for predicting the correct event only (Hurley and Shogren, 2005). This method is robust to deviations from risk neutrality and expected utility maximization. When the distribution is continuous one needs to elicit an interval, as explained below.

2.2.4 Eliciting quantiles and the median

One way to get a good idea of a cumulative distribution without eliciting the entire distribution is to elicit quantiles (Jose and Winkler, 2009). We call x a quantile α of the cdf F if $F(x) = \alpha$. Cervera and Muñoz (1996) presents a general scoring rule for the elicitation of quantile $\alpha \in (0,1)$, which is given by

$$S^{\alpha}(r,x) = \alpha r - (r-x)\mathbf{I}_{\{r \geq x\}}. \tag{6}$$

This rule rewards a high report, but punishes the subject if the report exceeds the realization, and is strictly proper for risk neutral subjects. Obviously, the median can be elicited by setting $\alpha = 0.5$.

2.2.5 Eliciting confidence intervals

Proper scoring rules exist for the elicitation of $\alpha{\cdot}100\%$ confidence intervals. Winkler and Murphy (1979) present a 'double' version of the quantile scoring rule discussed above, which elicits the $\frac{\alpha}{2}$ and $\frac{1+\alpha}{2}$ quantiles. The rule requires the subject to specify an upper bound u and a lower

[6] This mechanism can also be implemented by using a menu list of choices between a sure amount r and the prospect $y_E g$, where r is increasing for each choice. At the end, one decision is randomly selected for payment. It can also be presented as a scoring rule. Let $u(z)$ be the utility of prize z. Then $S(r,1) = P(Z \leq r)\, u(y) + \int_r^{\infty} u(z)\, dP_Z(z)$ and $S(r,0) = P(Z \leq r)\, u(g) + \int_r^{\infty} u(z)\, dP_Z(z)$ so that $ES(r,X) = P(Z \leq r)\,[pu(y) + (1-p)\, u(g)] + \int_r^{\infty} u(z)\, dP_Z(z)$.

bound l. Here we present an affinely transformed version of this rule which is given by

$$S^{Int}(l,u,x) = -\frac{(1-\alpha)}{2}(u-l) - (l-x)\,\mathbf{I}_{\{x\leq l\}} - (x-u)\,\mathbf{I}_{\{x\geq u\}}. \tag{7}$$

In words, this rule punishes the subject for specifying a larger interval width, and for the distance of x from the interval bound if x falls outside of the interval. Schmalensee (1976) presents a similar proper scoring rule that adds to (7) an extra term $\left|x - \frac{l+u}{2}\right|$ that penalizes the subject if the realization x is away from the mid-point of the interval. Schlag and van der Weele (2012) point out that these rules do not necessarily elicit the mode, nor the events that the subject thinks are most likely to occur. As a result they are 'imprecise', in that the chosen interval is often larger than necessary to cover $\alpha{\cdot}100\%$ of the mass (see also Section 2.7).

2.2.6 Eliciting continuous density functions

Matheson and Winkler (1976) shows that the quadratic, spherical and logarithmic scoring rules can be modified to generate proper scoring rules for density function elicitation. For example, the continuous quadratic scoring rule

$$S(r(x)) = 2r(x) - \int_{-\infty}^{\infty} r^2(x)dx, \tag{8}$$

is strictly proper for the density function.

An operational way to elicit density function is to discretize continuous distributions, and elicit probabilities for subsets of outcomes (Harrison *et al.*, 2013a). The elicitor may want to fit a distribution to these points ex-post (Garthwaite *et al.*, 2005). Another approach is to let the subject choose from a limited number of distributions the one which best approximates the true distribution. This implies some discrepancy between the true and elicited distribution. Friedman (1983) discusses the design of scoring rules to minimize this discrepancy.

2.3 Selecting between proper scoring rules

When multiple proper scoring rules exist for $\theta(X)$, the experimenter may apply selection criteria. These may be based on practical considerations. For example, she may select a rule that never involves payments from the subject to the experimenter, a rule with an upper bound on the possible payoffs, or one that is easier to explain to subjects (see Section 4).

From the perspective of decision theory, there are some arguments in favor of the QSR. Selten (1998) proves that the QSR obeys appealing axioms relating to the invariance of superficial changes to the elicitation environment. Here we present a new justification for the QSR, based on the strength of the incentives to tell the truth, which is a relevant criterion when the subject can exert effort to avoid mistakes or gather additional information.

Incentives to tell the truth can be measured by the curvature of S around the true report.

Since any affine transformation of a PSR produces a new PSR with different incentives, we first fix a range of payments $[\omega_1, \omega_2]$ and compare rules which have payments in this range. We use the characterization of Schervish (1989) to prove the following (the proof is in the Appendix).

Proposition 1 *Consider a scoring rule S for the elicitation of an event which has a reward in the interval $[\omega_1, \omega_2]$ and admits a piecewise continuous density in the Schervish representation. Consider the S^{QSR} as in (5) with $a = \omega_2$ and $b = \omega_2 - \omega_1$.*

If $S \neq S^{QSR}$, there exist p_0 and $\varepsilon > 0$ such that such that $\left|\frac{dES(r,X)}{dr}\right| < \left|\frac{dES^{QSR}(r,X)}{dr}\right|$ holds if $|r - p_0| < \varepsilon$ and $p_0 = EX$.

In words, the proposition states that the quadratic scoring rule will have stronger incentives for truth-telling than any other proper scoring rule in the neighborhood of some subjective probability. The result extends immediately to the elicitation of means as the probability of an event is a the mean of particular random variable.

Another selection criterion is simplicity. Here, one may favor the logarithmic scoring rule, which depends only on the probability given for the correct answer. It turns out that the logarithmic scoring rule is the unique proper scoring rule that has this property amongst differentiable scoring rules. McCarthy (1956) mentions this claim and attributes it to Gleason (unpublished). Since we were unable to locate the latter study, we prove this here for $n = 2$, using the framework of Savage (1971). We search for a rule such that $Y(r) = Z(1 - r)$ for all r, and hence $J'(p) = Y(p) - Z(p) = Y(p) - Y(1 - p)$. Since $J(p) = Y(p)\, p + Y(1 - p)(1 - p)$ we obtain $J'(p) = Y(p) - Y(1 - p) + Y'(p)\, p - Y'(1 - p)(1 - p)$ and hence $Y'(p)\, p = Y'(1 - p)(1 - p)$ for all p. This implies that $Y(p) = a \ln p + b$ for some $a > 0$ and b and hence S is an affine transformation of the the logarithmic payment scheme.

One may also also consider rules where the payments depends only on the report for the realized state, where these payments may be contingent on the state, i.e. $S(r, i) = f_i(r_i)$. An example of such a rule is the QSR where there are only two states. Shuford *et al.* (1966) prove that when $n > 2$, there are no unbounded rules in this class (as the logarithmic rule is unbounded). Thus, if there are more than two states, there are no rules with bounded payments that depend only on the probability reported for the right answer, and for $n = 2$ one needs to make the payments state contingent to obtain a bounded rule with this property.

2.4 Incentivized elicitation when subjects are not risk neutral

The proper scoring rules described above rely on assumptions about the preferences and the rationality of subjects. Perhaps the strongest of those assumption is that the subject is risk neutral. Countless laboratory studies have shown that most subjects behave as if they are risk averse over the stakes normally used in experiments (e.g. Holt and Laury, 2002). If subjects are not risk neutral, the scoring rules presented above are no longer proper (Winkler and Murphy, 1970). For example, under the QSR a risk averse subject should submit reports that are biased

away from extreme outcomes in order to minimize losses. Using the QSR, Offerman *et al.* (2009) and Armantier and Treich (2013) provide evidence that subjects do indeed report beliefs that are consistent with such a strategy.

What are the alternatives to assuming risk neutrality? First note that if the utility function of the subject is known, one can offset risk aversion by paying the subject in utils rather than money. Thus, $u^{-1}(S)$ is a proper scoring rule if S is a proper scoring rule for a risk neutral subject (Winkler, 1996). In the more typical case where the utility function is not known, Schlag and van der Weele (2013) show that truthtelling rules for the probability of an event with deterministic payments do not exist. In particular, none of the rules discussed in the previous subsection (including the QSR) are truth-telling. In this case, several strategies are open to the experimenter who does not want to assume risk neutrality.

2.4.1 Paying small stakes

Ramsey (1926) suggests to minimize distortions arising from risk preferences by paying small stakes. However, Armantier and Treich (2013) prove that this does not necessary solve the problem. Paying small stakes only reduces biases when subjects display increasing relative risk aversion, and worsens it for decreasing relative risk aversion. The authors find evidence for increasing relative risk aversion in an experiment, where biases found under elicitation with the QSR are significantly smaller when payoffs are low. Thus, paying small stakes for belief elicitation, which is the practice in economic experiments anyway, goes some way to addressing the problem. However, it may not eliminate the problem, as Holt and Laury (2002) finds substantial risk aversion even for low levels of incentives. Moreover, it may undermine the benefits of incentivized elicitation that motivated its use in the first place.

2.4.2 Randomized payments

Another option is to fix a single (monetary) prize, and let the scoring rule determine the probability of winning the prize. A subject with any risk preferences just wants to maximize the probability of getting the prize if they prefer the prize over getting nothing. Since expected utility is linear in probabilities, this procedure induces risk neutrality. We now discuss two implementations of this idea.

Paying in lottery tickets. One idea, due to Smith (1961)[7], and implemented in the context of belief elicitation by e.g. McKelvey and Page (1990), is to replace the deterministic rewards for an accurate guess with a probabilistic reward. Harrison *et al.* (2013b) and Hossain and Okui (2013) consider a lottery version of the QSR, which Hossain and Okui (2013) label the 'binarized scoring rule' (see Table 1 and Section 4.3). Schlag and van der Weele (2013) show

[7]There seems to be some confusion about the origin of this idea. Smith (1961) says the idea is 'adapted from Savage (1954)', but Savage (1971) attributes the idea to Smith.

how to generally apply the randomization of payoffs to the deterministic scoring rules discussed above. By appropriately normalizing the probability of winning the prize, all the proper scoring rules discussed in Section 2.2 that have bounded payoffs (so excluding the logarithmic rule) can be transformed into randomized rules that are truthtelling for all risk preferences.[8] For instance, assume that $S \in [\omega_1, \omega_2]$. Then one can give the prize to the subject with probability $\frac{S(r,x)-\omega_1}{\omega_2-\omega_1}$. In case $\omega_1 = 0$ it is as if one gives the subject S lottery tickets from a total set of ω_2 lottery tickets.

Reservation probabilities. Another variation of the reservation-price elicitation mechanism of Becker *et al.* (BDM, 1964) is to elicit reservation probabilities. The mechanism appears to have been invented by Ducharme and Donnell (1973) and variations have been proposed by Grether (1981), Allen (1987), Holt (2006) and Karni (2009). As in the previous mechanism, the report determines the probability of winning a fixed prize, inducing risk neutrality. Note that this method can also be represented as a scoring rule with randomized payoffs, a representation we omit here for reasons of space.

The experimenter asks the subjects to report the lowest probability r such that she is indifferent between a prospect y_r0 (i.e. a lottery which pays y with probability r and 0 with probability $1-r$) and the prospect y_E0. For payment, a number q is chosen according to a random variable Q that has distribution P_Q with support on $[0,1]$. The subject receives y_q0 if $q > r$ and y_E0 otherwise. Reporting the true subjective probability p maximizes expected utility: reporting $r < p$ leads to lower (expected) payoffs when q falls in $[r,p)$ and reporting $r > p$ backfires when q is in $(p,r]$.

This method can also be implemented using a menu list. Subjects choose multiple times between y_E0 and y_a0 where the value of a is gradually increased. When the subject indicated all her choices, one is randomly selected for payment. The value of a where the subject switches between lotteries is equal to the true subjective probability.[9]

2.4.3 Estimating deviations from risk neutrality

The methods in the previous subsection rely on theoretical assumptions about the ability to induce risk neutrality. Alternatively, the experimenter can estimate deviations from risk neutrality (and expected utility maximization, see below) on the basis of additional reports, and use these estimates to correct the elicited beliefs. This is the strategy proposed by Offerman *et al.* (2009) and Andersen *et al.* (2013).

[8] All definitions above immediately extend to randomized payment schemes, where the payment to the subject is a realization of some random variable. Here $S : \Theta \times \mathcal{X} \to \Delta\mathbb{R}$ where $\Delta\mathbb{R}$ denotes the set of distributions over $\mathbb{R}$.

[9] A problem arises when subjects do not have an unique switching threshold. Heinemann *et al.* (2009) excludes such subjects.

Offerman *et al.* (2009) propose to first elicit beliefs about events with known objective probability p. The elicited report function $R(p)$ shows how reports are biased away from the true beliefs due to deviations from risk neutrality. Subsequent reports r about other events with unknown subjective probability, can then be matched to the identical report $R(p)$, for which the underlying objective belief is known, and the true subjective belief p is recovered by inverting R, i.e. $p = R^{-1}(r)$.

To be able to match two identical probabilities, the experimenter needs to elicit (or otherwise be able to approximate) the correction function R^{-1} for the relevant range of p. Indeed, using the QSR to elicit beliefs about the throws of two 10 sided dice, Offerman *et al.* (2009) approximate the correction function, and provide evidence for substantial deviations from risk neutrality (and expected utility, see below). Note that this method does not require any structural assumptions or estimations, but uses actual reports to correct for risk aversion. Because of the substantial investment required to obtain correction function, the authors argue that the method is most attractive if subjective beliefs are elicited about a substantial number of events. By contrast, when only few beliefs are elicited the use of randomized payoffs (as explained above) may be preferable.

Andersen *et al.* (2013) present subjects with a range of bets and use maximum likelihood to jointly estimate the risk preferences and subjective beliefs, assuming a structural form for the subject's utility function and the decision making model. For this method, a considerable amount of data is required to estimate choices with some degree of confidence, which is time consuming to collect. A less resource intensive approach is to estimate an average subject's beliefs and utility functions, where these can be conditioned on background characteristics like gender and age.

2.5 Scoring rules for non-expected utility maximizers

Several authors have explored the possibilities for belief elicitation under more general assumptions about the decision making process of the subject. These models mostly assume 'probabilistic sophistication', which allows for probability weighting, but retains the assumption that beliefs are a probability measure. Harrison *et al.* (2013b) and Hossain and Okui (2013) show that the lottery version of the QSR (see Section 2.4.2) elicits truthfully in the context of an rank-dependent utility model. Andersen *et al.* (2013) estimate the subjective probabilities assuming a similar model, using the procedure described in the previous section.

Offerman *et al.* (2009) show that the correction procedure to the QSR discussed in Section 2.4.3 can also recover true subjective beliefs under probabilistic sophistication. However, they find that many subjects report a probability of 0.5 for a range of intermediate objective probabilities. This means the correction curve is not invertible, and subjective beliefs cannot be recovered accurately for reports of 0.5. To address this problem, Offerman and Palley (2013) consider the QSR when subjects are loss averse and form endogenous reference points. They show that loss

aversion can explain the over-reporting of 0.5, and provide a version of the QSR that corrects for this by underweighting payoffs that are perceived as losses. In an experiment, they find that this scoring rule leads subjects to correctly reproduce induced objective probabilities, obviating the need to elicit a correction function.

Kothiyal *et al.* (2011) also investigate the over-reporting of 0.5, and consider the performance of a more general set of scoring rules under probabilistic sophistication allowing for non-additive beliefs. The authors explain the bunching as the result of the reversal of payoff ranks at a belief of 0.5 and propose a *comonotonic* scoring rule that preserves the rank order of payoffs.

2.6 Other scoring rules and mechanisms

The linear scoring rule. The linear scoring rule has the following score when event j occurs

$$S^{Lin}(\mathbf{r}, x) = r_j. \tag{9}$$

Note that if the subject is approximately risk neutral, she has an incentive to report a probability of 1 for the event she thinks is most likely to occur, as this event has the highest marginal utility. However, if the subject has logarithmic utility one obtains the logarithmic scoring rule, which is strictly proper. Moreover, true subjective beliefs may be recovered if the utility function is known or estimated, although non-invertibility may occur like discussed in the previous section if subjects report 0 or 1 for a range of subjective beliefs.

Elicitation games. Perhaps the earliest elicitation mechanisms is the fair betting game, proposed by Toda (1951, see also Vlek (1973a)). The first player proposes a distribution of the total stake over two sides of a bet on the outcome of an uncertain binary event. The other player then chooses which side of the bet to take. In order to avoid ending up with the inferior side of the bet, it is optimal for the first player to make both bets equally attractive to the second player. Specifically, if the first player is risk neutral and believes that the second player has the same subjective probability about the outcome of the event, the proportional distribution according reflects her true belief p. Perhaps these two, rather strong assumptions explain the limited use of this mechanism in the literature (although see Vlek, 1973b).

Several papers consider elicitation where scoring is based on the reports of others instead of the realization of a random variable. These papers derive truthful elicitation as a Bayesian Nash equilibrium where it is optimal for each subject to report truthfully, as long as others do so. Prelec (2004) proposes to elicit both a subjective belief and an expectation of the frequency of all beliefs in the population. The subjective belief receives a score that is proportional to the difference between its actual frequency and the frequency predicted by the subject. Since a rational (Bayesian) person expects others to underestimate the prevalence of her own belief, it is optimal to report truthfully.

Miller *et al.* (2005) propose a mechanism where each subject is asked to predict the report of another subject after having received a signal about the state of the world. The elicitor, who is assumed to know the common prior, applies a proper scoring rule to the posterior probabilities of the subject that are implied by the subject's report. If signals about the state are correlated across subjects, truthful reporting is a strict (but not unique) equilibrium.

Prediction markets. Recent work has looked at popular (online) betting schemes known as prediction markets. In such markets, people trade claims that pay conditional on the occurrence of some (often political) event. A belief in efficient markets could lead one to think that the market price reflects the average belief in the market. However, Manski (2006), assuming risk neutral traders who take prices as given, shows that market prices do not pin down mean beliefs, although they do put a bound on it. Wolfers and Zitzewitz (2006) show that markets do reflect mean beliefs under particular assumptions about risk aversion and independence between wealth and beliefs. Using simulations, Fountain and Harrison (2011) show that prediction markets will not generally reflect mean prices when wealth, discount rates or risk aversion are correlated with beliefs.

2.7 Extensions

Here we discuss some extensions to the standard framework that we believe are promising avenues for further research.

Eliciting sets. When truth-telling schemes fail to exist, or are deemed too complicated for implementation, one may wish to elicit a set that contains the parameter of interest. The experimenter then asks the subject to report a set A that contains the parameter of interest θ. To simplify notation, consider the case where the parameter of interest is a real number, so $\Theta \subseteq \mathbb{R}$. Formally, the scheme $S = S(A, x)$ depends on the reported set $A \subset \Theta$ of the subject and on the realization x of X. In this case we say that S is *compatible with the truth* if

$$\theta(X) \in \arg\max_{A \subset \Theta} Eu(S(A, X)) \text{ for all } u \in U \text{ and all } P_X \in \mathcal{P}_X.$$

In the context of eliciting a set it may be simpler for implementation not to focus on truth telling schemes but instead to take the misreporting into account as follows. One asks the subject to report the parameter and then derives all parameters θ that could lead to this report for some utility function u. Formally, for any given scheme $S = S(r, x)$ and any report r one defines $A(r)$ by

$$\bar{\theta} \in A(r) \text{ if and only if } \bar{\theta} \in \arg\max_{r \in \Theta} Eu(S(r, X)) \text{ for some } u \in U \text{ and some } P_X \in \mathcal{P}_X \text{ with } \theta(X) = \bar{\theta}.$$

The set $A(r)$ is then the inference about θ one obtains from scheme S.

An example of this approach is Schlag and van der Weele (2012). The authors consider a random variable with support on $[a, b]$ and ask the subject to state an interval $[l, u]$. The score is given by

$$S^T(l,u,x) = \begin{cases} \left(1 - \left(\frac{l-u}{b-a}\right)\right)^{\frac{1-\alpha}{\alpha}} & \text{if } x \in [l,u] \text{ and } \frac{l-u}{b-a} \leq \alpha, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The rule rewards the subject if x is in the stated interval, where the reward declines in the width of the interval. The authors show that if the subject is (weakly) risk averse, the optimal interval contains the mode as well as a $\alpha{\cdot}100\%$ confidence interval for a realization of X. In addition, the rule truncates the reward when the interval is wider than than necessary to cover $\alpha{\cdot}100\%$ of the mass, improving the precision of the rule relative to other scoring rules for confidence intervals discussed in Section 2.2.5.

Paying on the basis of multiple events. One may wish to increase the number of independent events that are used to pay the subject. Sometimes this is necessary to elicit particular objects. For instance, one needs two independent realizations x_1 and x_2 of X, to elicit the variance of X when the subject is risk neutral (Lambert *et al.*, 2008). The payment scheme is given by $S(r, x_1, x_2) = a - \left(r - \frac{1}{2}(x_1 - x_2)^2\right)^2$ (see Schlag and van der Weele, 2013).

At other times, multiple events occur naturally in an experiment, for example when there is a group of people independently making the same choice. Hurley and Shogren (2005) provides a mechanism that asks for the empirical frequency of these choices. They show that from this report, the experimenter can recover an interval that contains the true subjective probability for each individual event. These intervals can achieve reasonable precision. For example, if one can pay based 20 realizations then one can derive an interval of less than 0.05 containing the true belief.

Multiple reports. One may also choose to increase the number of reports. For instance, one can elicit the variance when the subject is risk neutral by asking for two reports. Report r_1 is used to elicit the expected value of X^2, report r_2 to elicit the expected value of X. Then $r_1 - (r_2)^2$ elicits the variance of X if the subject is risk neutral. Note that EX^2 can be elicited when the subject is risk neutral by applying the quadratic scoring rule to x^2, so using $S(r_1, x) = a - \left(r_1 - x^2\right)^2$.

3 Comparing elicitation procedures

In the previous section we surveyed incentive schemes for eliciting beliefs. Here we review the literature that has compared the empirical performance of these schemes. We also include comparisons with what we label "introspection", i.e. simply asking for a probability or parameter of interest without the use of an incentive compatible elicitation method. Introspection may or may not be rewarded with a flat fee.

In many studies incentive compatible mechanisms are implemented in an unincentivized way by using hypothetical payoffs. Vlek (1973b) points out that even with hypothetical payoffs these mechanisms may still matter because they encourage subjects to think in a particular way and may align the preferences of experimenter and subject. One example of such a situation is in the elicitation of confidence intervals. Yaniv and Foster (1995, 1997) show that subjects seem to think that 50% confidence intervals strike the right balance between accuracy and preciseness even when the requested level of confidence is much larger.[10] An appropriate feedback rule would clarify what the experimenter really wants to know and avoid wrong interpretations. Winkler and Murphy (1968) provides a discussion of scoring rules as learning devices.

The main challenge in evaluating the performance of elicitation mechanism is the fact that the true belief, with which we wish to compare the elicited belief, is typically unobservable. Several approaches have been taken to deal with this problem. Perhaps the most prominent is to "induce" beliefs by informing subjects of the true probability, or giving them sufficient information to identify it. Three other benchmarks have been considered in the literature: how closely stated beliefs correspond with the empirical distribution; the degree to which they are consistent with behavior; and whether or not they satisfy additivity.[11] In Section 3.1 we discuss the advantages and disadvantages of each of these approaches. The references, methods and results of individual papers are summarized Table 1. In Section 3.2 we discuss the results and make suggestions for future work.

3.1 Methods of comparison

3.1.1 Inducing beliefs

The simplest way of inducing beliefs is by informing subjects directly about the objective probability of an event.[12] Hao and Houser (2012), for example, show subjects the number of black and white chips in a bag, and then try to recover the probability with which a randomly drawn chip is believed to be of a given color. It seems reasonable to assume that the subjective (i.e. induced) probability should be equal to the true probability, and that the smaller the distance between the latter and the elicited probability the better the elicitation method.

[10] This interpretation casts doubts on the widespread interpretation that intervals that are too wide are a sign of 'overconfidence'. Krawczyk (2011) shows that using incentives for interval elicitation improves the level of calibration of subjects.

[11] Two further techniques for testing the quality of belief elicitation methods have been considered in the literature. One possibility is to elicit beliefs about a number of events twice from the same subject and look at the correlation between the two responses. Of course this technique can only compare random errors associated with different methods and not systematic errors. The only such experiments we are aware of compare different response modes rather than methods and will be discussed in Section 4.4. Also, one can test that test that the reports generated by two different methods are the same, either within-subject (e.g. Beach and Phillips, 1967) or between-subject (e.g. Andersen *et al.* (2013) who find that the QSR and linear scoring rule produce similar reports after correcting for risk aversion and probability weighting). Such comparisons can show at best that at least one elicited is false. In this section we only consider papers that use a benchmark that is external to the elicited beliefs.

[12] Relevant studies are listed in Table 1, benchmark: Induced Probabilities (Direct).

There are two potential problems with this approach, one depending on context, the other on the elicitation methods in question. First, it is not clear that people respond to objective and subjective probabilities in the same way, especially if the subjective probability is something derived from a situation of strategic uncertainty. Second, this approach becomes trivial when comparing methods where the response must be in the form of a probability, such as introspection ("The probability A will occur is x. What is the probability A will occur?").

To address the second problem, a number of experimenters have attempted to induce beliefs by supplying theoretically sufficient information for subjects to calculate the true probability. This prevents subjects from simply repeating the probability of which they have just been informed. One commonly used method is to describe two distributions (e.g. two bingo cages with different proportions of red and white balls), show a series of draws with replacement from one of the two distributions, and elicit the subjects' posterior belief about the probabilities that the draws were from each of the distributions. The elicited beliefs can then be compared to probabilities calculated using Bayes' Law.[13] A second possibility is ask for the probability a particular combination of events will occur, where the probability of each individual event is known.[14]

TABLE 1 ABOUT HERE

A potential problem with these techniques is that subjects unfamiliar with probability theory are likely to use various heuristics to evaluate probabilities, possibly resulting in systematic differences between the true and induced probabilities. Such biases may augment or diminish biases resulting from the elicitation in unknown ways. A second problem is that they require computation on the part of the subjects and thus are likely to be sensitive to the use of incentives, whereas subjective probabilities in many environments of interest to economists are available to subjects intuitively and may not require financial rewards to induce sufficient effort to report accurately (Camerer and Hogarth, 1999).

3.1.2 Correspondence with the empirical distribution

A number of papers use actual outcomes or distributions of outcomes as a benchmark for comparing belief elicitation methods.[15] In experimental economics, the most commonly used procedure is to elicit beliefs about the action of another subject, then compare these probabilities to the empirical frequency in the treatment. Other options are to use distributions subjects should be somewhat familiar with (e.g. heights of males and females) or give them limited exposure to the distribution in question (e.g. show a bingo cage with balls of different colors so the precise numbers can only be estimated).

[13] See Table 1, benchmark: Induced Probabilities (Bayes' Rule).
[14] See Table 1, benchmark: Induced Probabilities (Multiple Events).
[15] See Table 1, benchmark: Empirical Distribution.

Of course there is no reason that subjects' beliefs should be correct. In fact, the idea that beliefs may be incorrect is one of the main motivating factors for economists to develop reliable methods of belief elicitation (see Manski (2004) for a criticism of rational expectations). Badly calibrated beliefs could lead to erroneous conclusions about the accuracy of elicitation methods. For example, when comparing a standard QSR, and a QSR corrected for risk attitudes, overconfidence bias when predicting the actions of others (leading to more extreme probabilities) would counteract the bias caused by a failure to account for risk aversion in the QSR, resulting in beliefs which are closer to empirical distribution but presumably further from subjective beliefs. Obviously, this criticism does not apply if the purpose of the elicitation is to obtain accurate predictions rather than obtaining the best measure of subjects' beliefs.

3.1.3 Consistency with behavior

Another benchmark that has been considered is the degree to which elicited beliefs are consistent with actions in games.[16] The idea is that under the assumption that subjective probabilities are crucial in determining choices, as is the case with most decision theories in economics, a stronger relationship between stated beliefs and choices indicates higher quality elicitation. This approach provides a natural testing ground because the relationship between beliefs and actions is precisely the context in which economists are often most interested in beliefs.

In order to check whether beliefs are in fact consistent with choices one must assume a model determining the relationship between beliefs and actions: the finding that elicited beliefs are not consistent with best-response behavior could be explained either by a failure of the elicitation method or the assumption that subjects best respond to their beliefs. Moreover, if a standard economic decision making model is assumed, one must know the utility function of subjects. We are not aware of a paper comparing elicitation methods in this way that has made a serious attempt to measure for example risk or social preferences.[17]

3.1.4 Additivity

A final option is comparing how close beliefs elicited using different methods are to satisfying additivity, i.e the condition that the probability of the union of mutually exclusive events is equal to the sum of the underlying probabilities. This implies for instance that the sum of the probabilities of exhaustive and mutually exclusive events add to one.[18] If subjective beliefs are truly additive, then is a necessary (although not sufficient) condition for a valid elicitation method is that stated beliefs satisfy additivity.

[16] See Table 1, benchmark: Consistency.

[17] Trautmann and van de Kuilen (2011), however, do compare consistency comparisons on the basis of three alternative utility functions: expected value, CRRA, and Fehr-Schmidt preferences.

[18] See Table 1, benchmark: Additivity.

3.2 Discussion of results

A simple overview of the existing studies, considering all approaches of comparison, gives little clue as to which method of elicitation is preferable as many results are contradictory. For example, Wang (2011) finds that the QSR results in stated beliefs closer to the empirical distribution than introspection, whereas Hollard *et al.* (2010) find the opposite, and Phillips and Edwards (1966)'s finding that the linear scoring rule is preferable to the QSR is contradicted by Palfrey and Wang (2009).

Opposing results could be due to many factors: the reliability of belief elicitation methods may depend on the domain in question (e.g objective/subjective probabilities); more complex methods may perform differently depending on the mathematical literacy of the subjects; different subject pools may have different risk attitudes; incentives may be more important for situations where there is a strong reason to misreport. This points to the necessity of comparing elicitation methods in the domain and with the subject pool for which they are to be used. They should also be rated based on criteria relevant to the purpose of the elicitation.

One strong conclusion can, however, be drawn: where stated beliefs do differ significantly between methods this tends to be in a way that is theoretically consistent with the presence of risk-averse subjects. Jensen and Peterson (1973) and Armantier and Treich (2013) find that steeper incentives lead to less extreme reported probabilities, and Trautmann and van de Kuilen (2011), Offerman *et al.* (2009); Offerman and Palley (2013) and Harrison *et al.* (2012) all find evidence that correcting the QSR for risk aversion improves performance. This should come as little surprise as it has been long established that risk aversion is prevalent in typical subject pools (in fact, Seghers *et al.* (1973) called into question the validity of PSRs for this very reason).

Distortions in stated beliefs will not always be a problem, for example if one is only interested in establishing a difference in distributions of beliefs in two populations with identical distributions of risk preferences. In most cases, however, it appears that methods that are robust to risk preferences are to be preferred. This is especially crucial if a variable of interest is correlated with risk aversion, such as testing gender differences in beliefs.

Another tentative pattern is that the relationship between beliefs and actions is stronger when incentives are used (Trautmann and van de Kuilen, 2011; Gächter and Renner, 2010). Hoffmann (2013) finds that subjects use dominated strategies less often when beliefs are elicited. These results are in line with the idea that incentives cause subjects to think harder or more systematically about the game, but more evidence is needed before any firm conclusions can be drawn.[19]

[19] An additional difficulty in drawing these conclusions is that they involve implicit assumptions about preferences. For example, in the context of public goods games, a deeper understanding of the game may have very different implications for selfish individuals (who would reduce contributions) or altruistic individuals (who would increase contributions). Indeed, in the public good game of Gächter and Renner (2010), the interpretation that elicitation improves understanding rests on the assumption that people are conditional cooperators. Note that in this study, the statistical effect is weak and the results are also consistent with a consensus effect or the use of stated beliefs to justify (selfish) actions. Note that the possibility of different distributions of social preferences

With respect to methodology, all the common approaches to comparing belief elicitation methods have serious drawbacks. We consider directly inducing probabilities the least problematic, the other benchmarks assume away the very phenomena that we are most interested in when studying subjective beliefs (miss-calibration, overconfidence, bounded rationality). Reliable recovery of induced probabilities should be considered a necessary, but perhaps not sufficient, condition for a good mechanism.

Given the fundamental impossibility of using the only inarguably appropriate benchmark, i.e. the true subjective belief, we suggest that the best approach is to take the theory seriously and test what is empirically verifiable. First of all, one can test the assumptions on which the validity of a mechanism is based. Paraphrasing Staël von Holstein (1970) (who in turn drew on De Finetti (1965)), several explicit and implicit assumptions underlying elicitation mechanisms can be enumerated.

1. The method must be incentive compatible, i.e. all the assumptions from which incentive compatibility is theoretically derived must hold. Typically this simply means that subjects must have a particular utility function, and behave according to a particular decision theory.

2. Subjects must understand the implications of the incentive scheme.

3. Subjects must understand the correspondence between their own beliefs and the probabilities (numerical or graphical) into which they are to be translated.

There is a wealth of experimental evidence related to the first point, with many decision theories proposed and tested (see, for example, Harrison and Rutström (2009)). Point 2 can also be tested in the lab, and often is to some extent in the form of comprehension pre-tests. Point 3 has been studied in the psychology and medical literature. These last two points will be discussed in the next section. If we can ascertain that all the assumptions necessary for a method to reveal true subjective probabilities have been met, comparing results to a contestable benchmark becomes less important.

4 Issues of implementation

4.1 Interactions between decisions and belief elicitation

Experimental economists are typically not interested only in beliefs for their own sake, but also their relationship with decisions, or using them in conjunction with decision data to help identify preferences or motivations. For these purposes it is necessary to elicit both decisions and beliefs from the same subjects and a decision must be made as to which to elicit first, or whether to

in different subject pools would thus reconcile some of the contradictions in the literature.

elicit them simultaneously. This raises two questions: does the elicitation of beliefs affect the decisions subjects make, and does the elicitation of decisions affect beliefs?[20]

TABLE 2 ABOUT HERE

The main reason put forward for belief elicitation potentially affecting decisions is that it may deepen subjects' understanding of a situation and make them act in a more sophisticated fashion. Conversely, choosing an action could influence elicited beliefs through several channels: a consensus bias (people assume others will act in the same way as themselves); justification to oneself or the experimenter that an action was morally acceptable by demonstrating a belief that one's action conformed to the norm (or increasing self-esteem by believing that an action was exceptional); a need to convince oneself that the correct action was chosen by holding beliefs that are consistent with that action; or a salience bias which makes the chosen action seem more probable.

Table 2 summarizes all the relevant papers we are aware of. The evidence is scanty and contradictory. Taking an action has been found to increase and decrease the accuracy of elicited beliefs. With a similar degree of inconclusiveness, belief elicitation is found to decrease, increase, and have no effect on contributions in public goods games. Erev *et al.* (1993) find that eliciting beliefs about the probability of events diverts attention from the size of payoffs and reduces expected value maximization. Guerra and Zizzo (2004), on the other hand, find no effect of belief elicitation on trusting behavior. Hoffmann (2013) compares an action-only treatment with a treatment where beliefs are elicited simultaneously, and finds that belief elicitation makes subjects less likely to choose dominated actions.

The small number of studies and apparently contradictory results on the two related methodological questions discussed in this section make it hard to draw any strong conclusion. It seems that eliciting beliefs can have an effect on decisions, but the direction of an effect, and the circumstances under which it arises is unclear. As discussed above, there is some evidence that belief elicitation affects play by deepening the understanding of the game. Overall, if independently measured beliefs and decisions are required from the same subjects, we can only recommend testing for an impact of belief elicitation on decisions and vice versa whenever designing a new game or using a new subject pool.

4.2 Hedging

Experimentalists are often interested in eliciting both decisions and probabilities from the same subject. However, paying subjects for both actions and elicitation tasks that depend on the

[20]Most of the literature discussed so far is based on the decision theoretic approach by Savage (1954), where subjective utilities are a primitive concept used in evaluating uncertain prospects. In contrast, psychologists have argued that choices may affect beliefs. A discussion of the merits of these approaches is beyond the scope of this paper and we limit ourselves discussing the empirical effect of *elicitation* on responses. Costa-Gomes *et al.* (2012) and Smith (2013) use an instrumental variable approach to identify a causal relationship between beliefs and actions.

outcome of the same event creates a situation in which subjects have a stake in the outcome of variable they are asked to predict. Kadane and Winkler (1988) and Karni and Safra (1995) show that under such circumstances proper scoring rules for the probability of an event no longer exist.[21] A specific concern in the context of experiments is that subjects have a incentive to hedge. For example, a risk averse subject facing the binary QSR who benefits when the event occurs, has an incentive to report an overly pessimistic belief in order to smooth her payoffs over the two states.

There is mixed evidence that hedging plays a role in economic experiments. Blanco *et al.* (2010) look at several games and contrast a hedge environment where both beliefs and choices are paid, with a no-hedge environment where only one of those is paid randomly. They find that a sizable number of subjects hedge in a 2×2 coordination game where the opportunity is obvious, but not in a more complex sequential prisoners' dilemma. The author's also find that in the former case, some subjects play a best response against other players' hedging strategies.

The results from Armantier and Treich (2013) confirm that subject may use obvious hedging opportunities. The authors used the QSR to elicit probabilities about events based on the roll of two dice. In a hedging treatment, subjects were also able to bet separately on the event in question. The authors find that subjects in the hedging condition bet more on the most likely events, and simultaneously report lower probabilities than in the control treatment. More circumstantial evidence comes from Palfrey and Wang (2009), which shows that observers with no stakes in the game and no incentive to hedge predict differently than subjects in the game.

Blanco *et al.* (2010) list a set of precautions that the experimenter can take to avoid hedging. First, one can elicit beliefs not about the matched partner's behavior, but about average behavior of the subjects in the partner's role, or a particular non-matched subject in that role (Armantier and Treich, 2009). This reduces the correlation between payoffs from belief reports and outcomes of play, and reduces the value of the hedge. Second, one can decide to randomly pay either the reported belief *or* the payoffs obtained in the game. Third, one may not pay for elicitation at all, although this may aggravate other sources of misreporting. Fourth, Blanco *et al.* (2010) find that some subjects hedge when they should not, so it may be helpful to explain subjects when they should not hedge. Finally, post-experiment questionnaires about the reasons for play and belief reports may also help detect hedging.

4.3 Complexity of incentive scheme

Many belief elicitation mechanisms require a high degree of mathematical sophistication (e.g. understanding the formulae of PSRs) or understanding relatively complex payment procedures (e.g. methods with probabilistic payoffs), and confusion among subjects has the potential to cause noise and bias in elicited beliefs (see Artinger *et al.*, 2010, for a discussion).

[21] Jaffray and Karni (1999) present mechanisms that can overcome these problems, which require either additional elicitation tasks, or the payment of very large sums of money to exploit the domain where the utility function is relatively flat.

In order to address the first problem, some experimenters present subjects with the formula for the PSR in question, assure them that stating their true belief will maximize the amount they can expect to earn, and offer a mathematical proof on request. A solution that is more in the spirit of revealed choice, is to have subjects select their preferred option from a list of bets generated using a scoring rule (e.g. Jensen and Peterson, 1973; Offerman *et al.*, 2009). With computers one can implement this easily by offering subjects sliders for setting the desired probability. When moving the sliders, the software can simultaneously display the payoffs associated with each outcome (e.g. Andersen *et al.* (2013)).

Whether or not subjects understand probabilistic payoffs schemes (Section 2.4.2) has been the subject of some debate (Berg *et al.*, 2008). Outside of a belief elicitation framework, Selten *et al.* (1999) casts doubt on the effectiveness of randomized payments in inducing risk neutrality. By contrast, in a very simple elicitation task with induced probabilities, Harrison *et al.* (2013b) and Hossain and Okui (2013) provide evidence that the use of probabilistic payoffs produces responses that are in line with risk neutral behavior, a finding which may or may not generalize to more complex environments.

4.4 Representation of probabilities

The format in which probabilities are communicated may matter, especially to subjects who are unfamiliar with them. Lipkus *et al.* (2001) find that people cannot in general convert between numerical probabilities, percentages, and frequencies, which suggests they are unlikely to respond the same way if asked for a subjective probability in different formats.

Some studies have addressed the question of which format subjects best understand. Tversky and Koehler (1994) who that find probabilities more likely to be additive if elicited as percentages rather than numerical probabilities. Gigerenzer and Hoffrage (1995) who find that subjects are better able to perform Bayesian updating when presented with frequencies rather than numerical probabilities. Price (1998) finds that eliciting probabilities as frequencies rather than numerical probabilities reduces the number of subjects expressing complete certainty, as well a measure of dispersion.

Probabilities can also be expressed graphically. Wang *et al.* (2002) find that the consistency of probabilities elicited at different times depends on whether they are elicited (from least to most consistent) as numbers, using a probability wheel, or a probability bar. Whitcomb *et al.* (1993) find no difference in consistency between elicitations as numbers, a probability wheel, or odds ratio.

4.5 Eliciting complementary events

It is common, especially with regard to binary events, to ask for the probabilities about all but one possible outcome, calculating the probability associated with the last by assuming additivity of subjective beliefs. Given the evidence that subjective beliefs appear to be consistently super-

additive (Tversky and Koehler, 1994; Trautmann and van de Kuilen, 2011), this is a questionable practice.

The finding of non-additivity can be a genuine feature of beliefs or an artifact of elicitation. If we believe that subjective beliefs are genuinely non-additive, we are forced to consider some theory that allows for this possibility (e.g. Gilboa, 1987).

Another explanation of super-additivity is that asking about a particular event increases its salience and makes it appear more likely, inflating the probabilities of each event. In this case, one possibility is to elicit probabilities for all events (i.e. A and "not A" for binary events) and deal with the resulting, inconsistent probabilities. Either the experimenter can scale them in some way to have them add to one, or the subject can reconcile the probabilities themselves. There is a substantial literature on the reconciliation of inconsistent probability assessments, e.g. Lindley *et al.* (1979). Alternatively the elicitation can be done in such a way that the input must be consistent, where care must be taken not to make one outcome more salient (e.g. order of elicitation). This can be achieved with the use of sliders.

5 Discussion and Conclusion

From personal conversations with colleagues, we have come away with the impression that opinions on the merits of using incentives for belief elicitation are divided. Roughly speaking, theorists or experimentalists with a strong theoretical focus tend to argue that incentivized elicitation is essential to interpret the elicited data. On the other hand, a sizable number of experimentalists are favorably disposed to non-incentivized elicitation and are comfortable to rely on intrinsic motivation of the subjects to answer questions correctly.

We believe that both sides have good arguments at their disposal, and that the relative strength of those arguments will depend on the context. Favoring the theorists, there are quite a few experimental situations where there are a priori reasons to assume that people may report falsely or sloppily. First of all, there are games in which subjects may use stated beliefs to justify their (selfish) behavior to the experimenter. This includes virtually all experiments which feature some trade-off between the payoffs of the decision maker and other subjects, such as dictator games, prisoner's dilemmas, public good games and trust games. Although we have not found direct evidence of such distortions, the evidence on the existence of experimenter demand effects (Zizzo, 2009) makes us believe that they should be taken seriously.

Second, subjects may simply 'click through' belief elicitation questions without putting in any effort, especially when they are bored or tired (at the end of a long experiment) and the questions are complex.[22] Third, there is some evidence that the use of scoring rules improves understanding of the game (Hoffmann, 2013) and consistency of decision making (Trautmann and van de Kuilen, 2011). Incentives may thus reduce noise in experimental data, although

[22] In Offerman *et al.* (1996), 50% of the subjects indicate that they would have reported different beliefs in the absence of incentives, often deviating to an 'easier' report.

the appropriate definition of 'noise' may depend on the aim of the study. A final reason to use scoring rules, but not necessarily incentives, is to clarify what is being elicited (see the discussion on overconfidence at the beginning of Section 3).

Balanced against these considerations are first and foremost are the practical costs of implementing and explaining incentive schemes. Trautmann and van de Kuilen (2011) compare the efforts required for different elicitation mechanisms in a table and show that these can be quite substantial. A second argument is that incentivization may create new distortions due to risk aversion or hedging. Note that there are trade-offs between these two arguments, as distortions due to risk aversion may be reduced by eliciting additional (and costly) reports.

In keeping with the above, unincentivized elicitation may be most advisable in situations where subjects are fresh, have no clear incentive to misreport, and face a straightforward elicitation task where the marginal benefit of subjects' effort is low and hedging may be a problem otherwise.[23] By contrast, using incentives is advisable in more complex or tedious tasks and when ruling out misunderstanding or careless reporting is especially important. An example of the latter are experiments testing cognitive biases. Engelmann and Strobel (2000), using incentivized belief elicitation, cannot reproduce the 'false-consensus effect' found in psychological studies that do not incentivize elicitation.

Suppose a researcher wishes to incentivize belief elicitation, which methods should she use? In answering this question we gather arguments from our discussion in the previous sections. One relatively clear result from the empirical literature is that risk aversion will bias beliefs elicited with proper scoring rules away from extreme probabilities. We would therefore recommend the use of corrective calibrations for such deviations, or the implementation of a randomized mechanism like reservation probabilities.

There are are several things the experimenter can do to ease the cognitive strain on subjects and encourage consistent reports. First, the use of sliders is beneficial for several reasons: it obviates the need for displaying complex formulae; no mention of probabilities is required; additivity is ensured and each event is given equal prominence. Second, when multiple observations are available, e.g. when larger groups of subjects in the session make the same decision, one can elicit the belief about the empirical frequency of a decision rather than the probability of a decision for a single person. Empirical frequencies seem to be more easily understood by subjects, and the procedure may also help to avoid hedging against the payoff-relevant decision of a single opponent.

What more do we need to know to conduct effective belief elicitation? The answer can be separated into a theoretical and empirical part. On the theoretical side, in Section 2.7 we have put forward extensions to the standard framework that await further elaboration. More

[23]Note that these conditions apply to most studies testing incentivized elicitation schemes, where belief elicitation is typically the only experimental task and thus receives full attention of the subjects. Therefore, these studies may understate effects of incentives in other, more complex, experimental settings (see the comments in Sonnemans and Offerman, 2001).

generally, an area of research that has seen considerable activity in recent years is the use of methods that are robust to different assumptions on preferences (e.g. risk aversion) and rationality (e.g. loss aversion). Given the evidence of heterogeneity in risk preferences and cognitive capacities, we believe this line of research should continue.

A second area concerns the trade-off between simplicity and informativeness. Given the limited time and resources that we can invest in belief elicitation, simple mechanisms that yield more imprecise information, for example by specifying bounds on beliefs, may be preferable to more complicated ones that yield very precise beliefs. A method that embodies advances in both these fields is the elicitation of intervals around empirical frequencies as in Hurley and Shogren (2005), elaborated in Section 2.7. This method is theoretically sound, robust to risk aversion, and is transparent to subjects as it only requires natural frequencies and involves a very simple payment rule.

Ultimately, the benefits of different incentivization mechanisms should be determined by empirical evidence, of which there is too little at present to draw any but tentative conclusions. Above, we have emphasized research investigating the validity of assumptions underlying different elicitation mechanisms. This includes both fundamental research on the nature of subjective beliefs and the feasibility of inducing a particular objective function in experimental settings.

More concretely, we need to improve our understanding about the interactions between belief elicitation and game play. If eliciting both choices and beliefs, one should experiment with the order of elicitation or even the presence of belief elicitation mechanisms in at least some sessions. There is a public good aspect to this kind of research, as it will help future researchers to make more informed design choices. Another important question is the separation of the cognitive effect of elicitation mechanisms (i.e. scoring rules as learning devices) and the incentive effects. To this end, we recommend that researchers testing incentive schemes include treatments that implement the incentive scheme with hypothetical payoffs. Finally, it would be valuable to test for the importance of experimenter demand effects, or justification of behavior through stated beliefs.

We hope that this paper will help experimentalists to make informed design choices and will provide inspiration for the development of new belief elicitation tools.

References

Allen, F. (1987). Discovering Personal Probabilities When Utility Functions Are Unknown. *Management Science*, **33** (4), 542–544.

Andersen, S., Fountain, J., Harrison, G. W. and Rutström, E. E. (2013). Estimating Subjective Probabilities. *Journal of Risk and Uncertainty*, **In Press**.

Armantier, O. and Treich, N. (2009). Subjective Probabilities in Games: A Solution to the Overbidding Puzzle. *International Economic Review*, **50** (4), 1079–1102.

— and — (2013). Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. *European Economic Review*, **In press**.

Artinger, F., Exadaktylos, F., Koppel, H. and Sääksvuori, L. (2010). *Applying Quadratic Scoring Rule transparently in multiple choice settings: a note*. Tech. rep., Jena Economic Research Paper.

Beach, L. and Phillips, L. (1967). Subjective Probabilities Inferred from Estimates and Bets. *Journal of Experimental Psychology*, **75** (3), 354–359.

Beach, L. R. and Wise, J. A. (1969). Subjective Probability Revision and Subsequent Decisions. *Journal of Experimental Psychology*, **81** (3), 561–565.

Becker, G., DeGroot, M. and Marschak, J. (1964). Measuring Utility by a Single-Response Sequential Method. *Behavioral science*, **9** (3), 226.

Berg, J. E., Rietz, T. A. and Dickhaut, J. W. (2008). On the Performance of the Lottery Procedure for Controlling Risk Preferences. *Handbook of Experimental Economics Results*, **1**, 1087–1097.

Blanco, M., Engelmann, D., Koch, A. K. and Normann, H.-T. (2010). Belief elicitation in experiments: is there a hedging problem? *Experimental Economics*, **13** (4), 412–438.

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

Camerer, C. C. F. and Hogarth, R. R. M. (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of risk and uncertainty*, **19** (1-3), 7–42.

Cervera, J. L. and Muñoz, J. (1996). Proper Scoring Rules for Fractiles. In J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), *Bayesian Statistics 5*, Oxford, U.K.: Oxford University Press, pp. 513–519.

Costa-Gomes, M. A., Huck, S. and Weizsacker, G. (2012). Beliefs and actions in the trust game: creating instrumental variables to estimate the causal effect. *WZB Discussion Paper*, **2012-302**.

Croson, R. T. A. (2000). Thinking like a game theorist: factors affecting the frequency of equilibrium play. *Journal of Economic Behavior & Organization*, **41** (3), 299–314.

Dawes, R. M., McTavish, J. and Shaklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *Journal of Personality and Social Psychology*, **35** (1), 1.

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, **18** (1), 87–123.

De Finetti, B. (1970). Logical foundations and measurement of subjective probability. *Acta Psychologica*, **34**, 129–145.

De Finetti, B. (1974). *Theory of Probability, Vol. 1*. New York: Wiley.

Delavande, A., Giné, X. and McKenzie, D. (2011). Measuring subjective expectations in developing countries: A critical review and new evidence. *Journal of Development Economics*, **94** (2), 151–163.

Ducharme, W. and Donnell, M. (1973). Intrasubject Comparison of Four Response Modes for "Subjective Probability" Assessment. *Organizational Behavior and Human Performance*, **10**, 108–117.

Engelmann, D. and Strobel, M. (2000). The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics*, **260** (2000), 241–260.

Erev, I., Bornstein, G. and Wallsten, T. (1993). The Negative Effect of Probability Assessments on Decision Quality. *Organizational Behavior and Human Decision Processes*, **55**, 78–94.

Fischer, G. W. (1982). Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. *Organizational Behavior and Human Performance*, **29** (3), 352–369.

Fountain, J. and Harrison, G. W. (2011). What do prediction markets predict? *Applied Economics Letters*, **18** (3), 267–272.

Friedman, D. (1983). Effective Scoring Rules for Probabilistic Forecasts. *Management Science*, **29** (4), 447–454.

Gächter, S. and Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, **13** (3), 364–377.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005). Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association*, **100** (470), 680–701.

Gigerenzer, G. and Hoffrage, U. (1995). How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological review*, **102** (4), 684–704.

Gilboa, I. (1987). Expected Utility with Purely Subjective Non-additive Probabilities. *Journal of Mathematical Economics*, **16**, 65–88.

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102** (477), 359–378.

Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B*, **14** (1), 107–114.

Grether, D. (1981). Financial Incentive Effects and Individual Decision-making. *California Institute of Technology, Working Paper 401.*

Guerra, G. and Zizzo, D. J. (2004). Trust responsiveness and beliefs. *Journal of Economic Behavior & Organization*, **55** (1), 25–30.

Hao, L. and Houser, D. (2012). Belief Elicitation in the Presence of Novice Participants: An Experimental Study. *Journal of Risk and Uncertainty*, **2** (April), 161–180.

Harrison, G. W., Martínez-Correa, J. and Swarthout, J. (2012). Eliciting Subjective Probabilities with Binary Lotteries. *Working Paper 2012-09, Center for the Economic Analysis of Risk, Robinson College of Business.*

—, Martínez-correa, J., Swarthout, J. T. and Ulm, E. R. (2013a). Scoring Rules for Subjective Probability Distributions. *Manuscript, Georgia State University.*

—, Martinez-Correa, J. and Swarthout, T. (2013b). Inducing Risk Neutral Preferences with Binary Lotteries: A Reconsideration. *Journal of Economic Behavior and Organization*, **94**, 145–159.

— and Rutström, E. E. (2009). Expected utility theory and prospect theory: One wedding and a decent funeral. *Experimental Economics*, **12** (2), 133–158.

Heinemann, F., Nagel, R. and Ockenfels, P. (2009). Measuring strategic uncertainty in coordination games. *Review of Economic Studies*, **76**, 181–221.

Hoffmann, T. (2013). The Effect of Belief Elicitation on Game Play. *Manuscript, Mannheim University*.

Hollard, G., Massoni, S. and Vergnaud, J. (2010). Subjective beliefs formation and elicitation rules : experimental evidence. *Centre d'Economie de la Sorbonne Working Paper*, **2010.88**.

Holt, C. (2006). *Markets, Games and Strategic Behavior*. Boston: Pearson/Addison-Wesley.

Holt, C. A. and Laury, S. (2002). Risk Aversion and Incentive Effects. *The American Economic Review*, **92** (5), 1644.

Hossain, T. and Okui, R. (2013). The Binarized Scoring Rule. *The Review of Economic Studies*, **In press**.

Huck, S. and Weizsäcker, G. (2002). Do players correctly estimate what others do?: Evidence of conservatism in beliefs. *Journal of Economic Behavior & Organization*, **47** (1), 71–85.

Hurley, T. and Shogren, J. (2005). An experimental comparison of induced and elicited beliefs. *Journal of Risk and Uncertainty*, **30** (2), 169—-188.

Hurley, T. T. M., Peterson, N. and Shogren, J. J. F. (2007). Belief Elicitation: An Experimental Comparison of Scoring Rule and Prediction Methods. *Manuscript, University of Minnesota*.

Jaffray, J. and Karni, E. (1999). Elicitation of subjective probabilities when the initial endowment is unobservable. *Journal of Risk and Uncertainty*, **8**, 5–20.

Jenkinson, D. (2005). The Elicitation of Probabilities - A Review of the Statistical Literature. *Manuscript, University of Sheffield*.

Jensen, F. A. and Peterson, C. R. (1973). Psychological effects of proper scoring rules. *Organizational Behavior and Human Performance*, **9** (2), 307–317.

Jose, V. R. R. and Winkler, R. L. (2009). Evaluating Quantile Assessments. *Operations Research*, **57** (5), 1287–1297.

Kadane, J. and Winkler, R. (1988). Separating probability elicitation from utilities. *Journal of the American Statistical Association*, **83** (402), 357–363.

Karni, E. (2009). A Mechanism for Eliciting Probabilities. *Econometrica*, **77** (2), 603–606.

— and Safra, Z. (1995). The Impossibility of Experimental Elicitation of Subjective Probabilities. *Theory and Decision*, **38**, 313–320.

Koessler, F., Noussair, C. and Ziegelmeyer, A. (2012). Information aggregation and belief elicitation in experimental parimutuel betting markets. *Journal of Economic Behavior & Organization*, **83** (2), 195–208.

Kothiyal, A., Spinu, V. and Wakker, P. (2011). Comonotonic Proper Scoring Rules to Measure Ambiguity and Subjective Beliefs. *Journal of Multi-Criteria Decision Analysis*, **113** (March), 101–113.

Krawczyk, M. (2011). Overconfident for real? Proper scoring for confidence intervals. *Manuscript, University of Warsaw*.

Lambert, N., Pennock, D. and Shoham, Y. (2008). Eliciting properties of probability distributions: the highlights. *ACM SIGecom Exchanges*, **7** (3), 1–5.

Lindley, D. V., Tversky, A. and Brown, R. V. (1979). On the Reconciliation of Probability Assessments. *Journal of the Royal Statistical Society. Series A (General)*, **142** (2), 146.

Lipkus, I., Samsa, G. and Rimer, B. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, **21**, 37–44.

Manski, C. (2002). Identification of decision rules in experiments on simple games of proposal and response. *European Economic Review*, **46**, 880–891.

— (2004). Measuring expectations. *Econometrica*, **72** (5), 1329–1376.

Manski, C. F. (2006). Interpreting the predictions of prediction markets. *Economics Letters*, **91** (3), 425–429.

Matheson, J. and Winkler, R. (1976). Scoring rules for continuous probability distributions. *Management Science*, **22** (10), 1087–1096.

McCarthy, J. (1956). Measures of the Value of Information. *Proceedings of the National Academy of Sciences of the United States of America*, **42** (9), 654–655.

McKelvey, R. and Page, T. (1990). Public and private information: An experimental study of information pooling. *Econometrica*, **58** (6), 1321–1339.

Miller, N., Resnick, P. and Zeckhauser, R. (2005). Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science*, **51** (9), 1359–1373.

Nyarko, Y. and Schotter, A. (2002). An Experimental Study of Belief Learning Using Elicited Beliefs. *Econometrica*, **70** (3), 971–1005.

Offerman, T. and Palley, A. B. (2013). Lossed in Translation : An Off-the-Shelf Method to Recover Probabilistic Beliefs from Loss-Averse Agents. *Manuscript, University of Amsterdam*.

—, Sonnemans, J. and Schram, A. (1996). Value Orientations, Expectations and Voluntary Contributions in Public Goods. *The Economic Journal*, **106** (437), 817–845.

—, —, Van de Kuilen, G. and Wakker, P. P. (2009). A Truth Serum for Non-Bayesians. *Review of Economic Studies*, **76** (4), 1461–1489.

Palfrey, T. R. and Wang, S. W. (2009). On eliciting beliefs in strategic games. *Journal of Economic Behavior & Organization*, **71** (2), 98–109.

Phillips, L. D. and Edwards, W. (1966). Conservatism in a Simple Probability Inference Task. *Journal of Experimental Psychology*, **72** (3), 346–354.

Prelec, D. (2004). A Bayesian Truth Serum for Subjective Data. *Science*, **306** (October), 462–466.

Price, P. (1998). Effects of a Relative-Frequency Elicitation Question on Likelihood Judgment Accuracy: The Case of External Correspondence. *Organizational behavior and human decision processes*, **76** (3), 277–297.

Ramsey, F. (1926). Truth and Probability. In R. B. Braithwaite (ed.), *The Foundations of Mathematics and other Logical Essays*, 1926, New York (1931): Harcourt, pp. 156–198.

Roby, T. B. (1964). Belief States: A Preliminary Empirical Study. *Technical Documentary Report, Decision Sciences Laboratory.*

Rutström, E. E. and Wilcox, N. T. (2009). Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. *Games and Economic Behavior*, **67** (2), 616–632.

Savage, L. J. (1954). *The Foundation of Statistics*. New York: Wiley.

— (1971). Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association*, **66** (336), 783–801.

Schervish, M. (1989). A General Method for Comparing Probability Assessors. *The Annals of Statistics*, **17** (4), 1856–1879.

Schlag, K. H. and van der Weele, J. J. (2012). Incentives for Interval Elicitation. *Manuscript Vienna University.*

— and — (2013). Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming Risk Neutrality. *Theoretical Economics Letters*, **03** (1), 38–42.

Schmalensee, R. (1976). An Experimental Study of Expectation Formation. *Econometrica*, **44** (1), 17–41.

Schum, D. A., Goldstein, I. L., Howell, W. C. and Southard, J. F. (1967). Subjective Probability Revisions under Several Cost-Payoff Arrangements. *Organizational Behavior and Human Performance*, **2**, 84–104.

Seghers, R. C., Fryback, D. G. and Goodman, B. C. (1973). *Relative Variance Preferences in a Choice-Among-Bets Paradigm*. Tech. rep., DTIC Document.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, **62**, 43–62.

—, Sadrieh, A. and Abbink, K. (1999). Money does not induce risk neutral behavior, but binary lotteries do even worse. *Theory and Decision*, **46**, 211–249.

Shuford, E., Albert, A. and Massengill, H. E. (1966). Admissible Probability Measurement Procedures. *Psychometrika*, **31** (2), 125–145.

Smith, A. (2013). Estimating the causal effect of beliefs on contributions in repeated public good games. *Experimental Economics*, **16** (3), 414–425.

Smith, C. (1961). Consistency in statistical inference and decision. *Journal of the Royal Statistical Society. Series B*, **23** (1), 1–37.

Sonnemans, J. and Offerman, T. T. T. (2001). Is the quadratic scoring rule behaviorally incentive compatible? *Manuscript, University of Amsterdam*.

Staël von Holstein, C.-A. S. (1970). Measurement of subjective probability. *Acta Psychologica*, **34**, 146–159.

Toda, M. (1951). Measurement of intuitive probability by a method of game. *Japanese Journal of Psychology*, **22**, 29–40.

— (1963). Measurement of Subjective Probability Distribution. *Report, State College, Pennsylvania, Institute for Research, Division of Mathematical Psychology*, **3** (3).

Trautmann, S. T. and van de Kuilen, G. (2011). Belief Elicitation: A Horse Race Among Truth Serums. *CENTER discussion paper 117, Tilburg University*.

Tversky, A. and Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychological Review*, **101** (4), 547.

Vlek, C. (1973a). The fair betting game as an admissible procedure for assessment of subjective probabilities. *British Journal of Mathematical and Statistical Psychology*, **26** (1), 18–30.

Vlek, C. A. J. C. (1973b). Coherence of human judgment in a limited probabilistic environment. *Organizational Behavior and Human Performance*, **9** (460-481), 460–481.

Wang, H., Dash, D. and Druzdzel, M. J. (2002). A method for evaluating elicitation schemes for probabilistic models. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **32** (1), 38–43.

Wang, S. W. (2011). Incentive effects: The case of belief elicitation from individuals in groups. *Economics Letters*, **111** (1), 30–33.

Whitcomb, K. M., Önkal, D., Benson, P. G. and Curley, S. P. (1993). An evaluation of the reliability of probability judgments across response modes and over time. *Journal of Behavioral Decision Making*, **6** (4), 283–296.

Wilcox, N. T. and Feltovich, N. (2000). Thinking like a game theorist: Comment. *University of Houston Department of Economics working paper*.

Winkler, R. (1996). Scoring rules and the evaluation of probabilities. *Test*, **5** (1), 1–60.

— and Murphy, A. (1968). "Good" Probability Assessors. *Journal of Applied Meteorology*, **7** (October), 751.

— and — (1970). Nonlinear utility and the probability score. *Journal of Applied Meteorology*, **9** (February), 143–148.

— and — (1979). The use of probabilities in forecasts of maximum and minimum temperatures. *The Meteorological Magazine*, **108** (1288), 317–329.

Wolfers, J. and Zitzewitz, E. (2006). Interpreting Prediction Market Prices as Probabilities. *NBER Working Paper 12200*.

Yaniv, I. and Foster, D. (1995). Graininess of Judgement Under Uncertainty: An Accuracy-Informativeness Trade-Off. *Journal of Experimental Psychology: General*, **124** (4), 424–432.

— and — (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, **10**, 21–32.

Zizzo, D. J. (2009). Experimenter demand effects in economic experiments. *Experimental Economics*, **13** (1), 75–98.

Table 1: Empirical comparison of belief elicitation mechanisms

List of abbreviations used:
BSR: binarised scoring rule (paying in lottery tickets, Section 2.4.2)
CE: certainty equivalent (Section 2.2.1)
LgSR: log scoring rule (Section 2.2.1)
LnSR: linear scoring rule (Section 2.6)
PSR: proper scoring rule (Section 2.2)
QSR: quadratic scoring rule (Section 2.2.1)
RP: reservation probability (Section 2.4.2)
SSR: spherical scoring rule (Section 2.2.1)

Reference	**Benchmark**	**Elicitation Method**	**Results**
Beach and Wise (1969)	Induced probability (Bayes' Rule)	CE (hypothetical payoffs), Introspection	No significant difference in means. Greater variance with CE.
Ducharme and Donnell (1973)	Induced probability (Bayes' Rule)	Best guess in session wins prize, RP	No significant difference.
Phillips and Edwards (1966)	Induced probability (Bayes' Rule)	Introspection, LgSR, LnSR, QSR	From closest to furthest from probability calculated by Bayes' Rule: LgSR, LnSR, QSR, Introspection.
Schum *et al.* (1967)	Induced probability (Bayes' Rule)	LgSR, LnSR	LgSR closer to Bayes' Rule than LnSR.
Sonnemans and Offerman (2001)	Induced probability (Bayes' Rule)	Introspection, QSR	No significant difference.
Hao and Houser (2012)	Induced probability (Direct)	Two versions of RP: declaritive and clock mechanisms.	Clock mechanism closer to true probability.
Hossain and Okui (2013)	Induced probability (Direct)	BSR, QSR	BSR closer to true probability.
Hurley *et al.* (2007)	Induced probability (Direct/Multiple events)	Prize for correct prediction, QSR	Less bias with QSR on average; best method is subject specific.

Continued on next page

Table 1 – continued from previous page

Reference	**Benchmark**	**Elicitation Method**	**Results**
Jensen and Peterson (1973)	Induced probability (Direct)	LgSR,QSR, SSR (plus two affine transformations of each)	No significant difference between different PSRs. Steeper incentives lead to more conservative probabilities. Rules with both +ve and -ve payoffs led to non-optimal strategies.
Offerman and Palley (2013)	Induced probability (Direct)	QSR, QSR corrected for loss aversion	Conservative bias under standard QSR. Corrected QSR reduces bias and leads to accurate reports.
Seghers *et al.* (1973)	Induced probability (Direct)	Unconventional PSRs: incentivised and hypothetical payoffs.	More accurate with hypothetical payoffs.
Armantier and Treich (2013)	Induced probability (Multiple events)	QSR: hypothetical, low and high stakes	From closest to furthest from true probability: hypothetical, low, high. Variance highest with hypothetical payoffs.
Fischer (1982)	Empirical distribution	Introspection, LgSR	No significant difference.
Gächter and Renner (2010)	Empirical distribution / Consistency	Non-incentive compatible scoring rule, Introspection	Scoring rule makes more accurate predictions and results in stronger relationship between beliefs and actions.
Harrison *et al.* (2012)	Empirical distribution	BSR, QSR, QSR (corrected for risk preferences)	QSR least accurate. No difference between the other two methods.
Hollard *et al.* (2010)	Empirical distribution	Introspection, QSR, RP	From most to least accurate: RP, introspection, QSR.
Huck and Weizsäcker (2002)	Empirical distribution	CE, QSR	QSR closer to true frequency.

Continued on next page

Table 1 – continued from previous page

Reference	**Benchmark**	**Elicitation Method**	**Results**
Krawczyk (2011)	Empirical distribution	Introspection, Interval Scoring Rule	Incentivized intervals are better calibrated although still too narrow.
Palfrey and Wang (2009)	Empirical distribution	LgSR, LnSR, QSR	LgSR and QSR better calibrated than LnSR. LnSR more extreme reports.
Trautmann and van de Kuilen (2011)	Empirical distribution / Consistency / Additivity	CE, CE (corrected for risk preferences), Introspection, QSR, QSR (corrected for risk preferences), RP	No significant differences in accuracy. Beliefs from introspection do not predict actions: all incentivised methods predict equally well. Additivity bias from most to least: QSR, QSR (corrected), introspection.
Vlek (1973a)	Empirical distribution	Fair Betting Game, Introspection	No significant difference.
Wang (2011)	Empirical distribution	Introspection, QSR	QSR more accurate, better calibrated, and more extreme reports.
Offerman *et al.* (2009)	Additivity	QSR, QSR (corrected for risk preferences)	Less bias after correction for risk preferences.

Table 2: Interactions between decisions and belief elicitation.

Reference	**Elicited first**	**Game/Decision**	**Results**
Erev *et al.* (1993)	Beliefs	Intergroup public goods / Choice of gambles on sport	Belief elicitation reduces expected value maximising behaviour.
Croson (2000)	Beliefs	Public goods / Prisoners' dilemma	Belief elicitation reduces contributions and cooperation.
Wilcox and Feltovich (2000)	Beliefs	Public goods	No significant difference.
Guerra and Zizzo (2004)	Beliefs	Trust game	No significant difference
Gächter and Renner (2010)	Beliefs	Public goods	Belief elicitation (sometimes) increases contributions.
Nyarko and Schotter (2002)	Beliefs	Asymmetric matching pennies	No significant difference.
Rutström and Wilcox (2009)	Beliefs	Asymmetric matching pennies	Difference only in early rounds and for only one player.
Koessler *et al.* (2012)	Beliefs/Decisions	Parimutuel betting market	Belief elicitation improves information aggregation. Betting improves accuracy of elicited beliefs.
Dawes *et al.* (1977)	Decisions	Prisoners' dilemma	Choosing action increases variance of beliefs.
Offerman *et al.* (1996)	Decisions	Public goods	No significant difference.
Palfrey and Wang (2009)	Decisions	Matching pennies.	Choosing action decreases accuracy of elicited beliefs.
Hoffmann (2013)	Simultaneous	Variety of normal form games	Eliciting beliefs decreases number of dominated strategies chosen.

Appendix

Proof of Proposition 1. We use the characterization of Schervish (1989). To simplify exposition assume that ν has no point masses and admits a piecewise continuous density f, hence $S(r,1) = S(1,1) - \int_r^1 (1-c) f(c)\, dc$ and $S(r,0) = S(0,0) - \int_0^r c f(c)\, dc$. Consequently, if $EX = p$ then

$$E(r,X) = pS(1,1) + (1-p) S(0,0) - p\int_r^1 (1-c) f(c)\, dc - (1-p)\int_0^r c f(c)\, dc$$

and

$$\frac{d}{dr} ES(r,X) = (p-r) f(r).$$

So $f(r)$ describes the strength of the local incentives to tell the truth for reports that are close to r.

Now note that

$$\begin{aligned} ES(1,X) - ES(0,X) &= -(1-p)\int_0^1 c f(c)\, dc + p\int_0^1 (1-c) f(c)\, dc \\ &= \int_0^1 (p-c) f(c)\, dc \\ &\leq \int_0^1 (1-c) f(c)\, dc. \end{aligned} \tag{11}$$

Assume now w.l.o.g. that the scoring rule gives payoffs in $[0,k]$ (i.e. $\omega_1 = 0$ and $\omega_2 = k$). For instance, the quadratic scoring rule would be represented as $S^{QSR}(r,1) = k\left(1-(1-r)^2\right)$ and $S^{QSR}(r,0) = k\left(1-r^2\right)$. It is easy to show that for the QSR, $f(r) = 2k$. Note that

$$ES(1,X) - ES(0,X) \leq k = 2k\int_0^1 (1-c)\, dc. \tag{12}$$

Comparing (11) and (12) it follows that $f \equiv 2k$ if $f(c) \geq 2k$ for all c. ■