Efficient Interval Scoring Rules

Karl H. Schlag* Joël van der Weele†

October 9, 2009

Abstract

Scoring rules that elicit an entire belief distribution through the elicitation of point beliefs are time-consuming and demand considerable cognitive effort. Moreover, the results are valid only when agents are risk-neutral or when one uses probabilistic rules. We investigate a class of rules in which the agent has to choose an interval and is rewarded (deterministically) on the basis of the chosen interval and the realization of the random variable. We formulate an efficiency criterion for such rules and present a specific interval scoring rule. For single-peaked beliefs, our rule gives information about both the location and the dispersion of the belief distribution. These results hold for all concave utility functions.

Keywords: Belief elicitation, scoring rules, subjective probabilities.

JEL Codes: C60, C91, D81.

1 Introduction

The subjective beliefs of agents are crucial determinants of behavior in many situations. Finding out these beliefs is thus an equally crucial matter not only for policy makers, but also for scientists who test models under maintained assumptions about beliefs. Mechanisms called *proper scoring rules* have been designed to elicit subjective beliefs or probabilities (Murphy and Winkler 1970, Gneiting and Raftery 2007). In such mechanisms the elicitor asks the agent to report a specific element or parameter of the belief distribution. She then rewards the agent in a way such that the agent has an incentive to report truthfully.

Unfortunately, the most popular of these scoring rules, like the well-known quadratic scoring rule or QSR (Brier 1950), suffer from some practical as well as theoretical drawbacks. These

*Universitat Pompeu Fabra, Barcelona. E-mail: karl.schlag@upf.edu. Karl Schlag greatfully acknowledges financial support from the Department of Economics and Business of the Universitat Pompeu Fabra, from the Universitat Pompeu Fabra, Grant AL 12207, and from the Spanish Ministerio de Educacion y Ciencia, Grant MEC-SEJ2006-09993.

†Goethe Universität, Frankfurt am Main. E-mail: vdweele@econ.uni-frankfurt.de

scoring rules elicit point beliefs, i.e. beliefs about a specific event or a specific characteristic of the distribution. To know the entire distribution therefore, the elicitor has to elicit probabilities over all possible events. From a practical point of view, if the support is large, this procedure is time consuming and cognitively demanding for the subject. The elicitor may therefore choose to gain an understanding of only part of the distribution such as the mean or the median, or must add parametric assumptions. A related problem is that eliciting beliefs with an intrusive method potentially changes the perception of the agents about the tasks that they are asked to perform after the elicitation. In the context of economic experiments, Rutström and Wilcox (2008) and Croson (2000) suggest that more intrusive elicitation mechanisms may change agent's strategic choices.

These considerations limit the use of QSR in the practice of economic experiments, in which the elicitation of beliefs is often only a small subset of the choices that the agents make and hence cannot take up too much time. As a consequence, researchers have sometimes abandoned the QSR or sought pragmatic tradeoffs between elicitation costs and informational content. For example, McDaniel and Rutström (2001) divide the support in separate intervals for which they then incentivize agents using a QSR. Gächter (2006) and Croson (2000) simply reward a small interval around a single correct prediction by the agents. Dufwenberg and Gneezy (2000) ask agents for a guess of the average value of expected play by others. Other researchers use the QSR to elicit only the expected value of the belief distribution (e.g. Huck and Weiszäcker, 2002). While these procedures minimize the cost of elicitation, they give only limited information about the distribution.

A second problem of deterministic elicitation methods, such as the QSR, is that these only lead the agent to report her true beliefs if she is risk neutral. This is obviously and important disadvantage, and the literature has yielded some ways around this. Offerman *et al.* (2009) use proper scoring rules like the QSR to correct reported probabilities for risk attitudes. To do so, one has to elicit the beliefs of a agent for a set of known objective probabilities p. However this remedy adds to the practical problems mentioned above, because it increases cost, time and cognitive complexity of the elicitation procedure. Fountain (2002) designs a log-scoring rule that elicits truthfully for the class of CARA-utility functions. It is possible to elicit point beliefs for more general risk preferences only if one resorts to randomized rules (Schlag and Van der Weele 2009), which in theory should lead the agent to be risk neutral. However, randomized rules are more difficult to implement, and some doubt has been raised about their effectiveness in inducing risk neutrality (Selten *et al.* 1999).

In this paper, we depart from the focus on proper scoring rules and the elicitation of point beliefs and consider a different class of rules, known as interval scoring rules. Such rules ask the agent to specify a subset of the support of a random variable X, and reward the agent on the basis of the chosen interval and a realization of X. We will show that interval scoring rules can provide an appealing trade-off between the cost of elicitation and the intrusiveness of the procedure on the one hand, and the information that is gained about the belief distribution on the other. Moreover, the scoring rule we propose allows inferences that are valid under any degree of risk aversion of the agent.

Since the rules we investigate do not ask the agent for a specific element of the distribution, the notion of 'propriety' or truthful-telling is not applicable. Therefore, we first develop an alternative measure to evaluate the performance of such rules. To do so we propose the notion *coverage*. By this we mean the minimal mass that an agent who maximizes expected utility will include in the interval, where the minimum is taken over all concave utility functions and all possible belief distributions. On the basis of this notion we formulate a criterion of *efficiency*, which says that a rule is efficient if there is not some other rule that has the same coverage but captures it in a smaller interval. Efficiency thus implies that the inferences on the distribution that can be drawn from the elicited interval are as precise as possible.

To show the applicability of interval scoring rules we the propose a specific scoring rule, that we call the *truncated interval scoring rule* (TISR). This rule rewards the agent if a realization of a random variable falls inside the specified interval and punishes the agent for specifying a wider interval. Under the assumption that beliefs are single-peaked, we are able to characterize the coverage of the TISR, and show that it is efficient. This means that the TISR can be used to locate an amount of mass that is bounded below, and that it does so in the most precise way possible. This result does not depend on risk preferences. Moreover, we can derive several additional features of the TISR that make inference on both the location and the dispersion of the rule possible.

Given the advantages of interval elicitation over point elicitation, it is perhaps surprising that interval scoring rules have received little attention. Although there is some literature on interval scoring rules, we are to our knowledge the first to posit a formal criterion to evaluate the performance of interval scoring rules. Some authors have studied the optimal choice of interval for a risk neutral agent under different interval scoring rules. Aitchison and Dunsmore (1968) study the behavior of optimal forecasting intervals under a limited class of differentiable densities. They show the first order conditions for the optimal interval, and give some applications to specific families of distributions such as the normal and exponential distributions. Winkler (1972) does a similar exercise for a more general class of interval scoring rules. To our knowledge however, no-one has systematically investigated the inferences an elicitor can make on the beliefs of the decision maker from an observed interval choice.

Nevertheless, different variants of interval scoring rules have been used in experiments. In his experiment on expectation formation, Schmalensee (1976) constructs an interval scoring rule that is close to the TISR, although with an extra penalty parameter. However, the inferences he makes are based on the assumption of risk neutrality, which we relax in this paper. In experiments on temperature forecasts, some authors have elicited intervals using scoring rules (e.g. Murphy and Winkler 1974). However, in most of these studies, the focus is either on 'credible intervals' with fixed probability (e.g. the forecaster is asked to specify an interval in which he believes the temperature to be with some probability, say 50%) or on intervals with a fixed width. Hamill and Wilks (1995) use a scoring rule to elicit an interval with variable width for minimum and maximum temperatures. Galbiati *et al.* (2009) use an untruncated version of the scoring rule proposed in this paper. In their study beliefs are elicited about the choices of another player in a minimum effort coordination game with large action spaces.

The paper proceeds as follows. In Section 2 we define the notion of *coverage* and use this to formulate a general criterion of efficiency for interval scoring rules. In Section 3 we define an interval scoring rule that we call the TISR. We characterize its coverage and show that the TISR satisfies the efficiency criterion developed in Section 2. In Section 4 we elaborate on the inferences that can be drawn under the TISR. All proofs are in the appendix.

2 Efficient Interval Scoring Rules

2.1 Preliminaries

Consider an agent endowed with preferences over $\mathbb{R}$ that admit an expected utility representation, denoted by u. This agent has subjective beliefs over the distribution of a random variable X that generates outcomes belonging to the state space Ω. The results of the paper are based on the assumption that this space is compact and known to be

$$\Omega = [a, b]\,. \tag{A1}$$

The set of all random variables that generate outcomes in Ω will be denoted by Δ. With respect to the preferences of the decision maker we assume that u is concave and continuously differentiable. Thus, our analysis includes both risk-neutral and risk averse agents.

To elicit characteristics of the beliefs about X, the elicitor (she) asks the agent (he) to specify boundaries L and U of the interval and pay the agent an amount $S(L, U, x)$ as a function of the realization x. The payment function S is also called a *scoring rule.* The set of all such scoring rules is denoted by $\mathcal{S}$.

On an intuitive level it is easy to understand how such interval scoring rules can be used to gain an idea of the distribution. If the agent is rewarded when he is 'correct', i.e. the realization x falls inside the interval, he will be tempted to specify the interval such that it contains a large probability mass. Moreover, if he is penalized for the width of the interval, he will face a trade-off that will lead him to exclude regions with low probability mass. The rest of this paper is dedicated to making this reasoning more precise.

Two main points of interest to the elicitor are the width of the interval and the mass that is contained within the interval. The combination of these two variables allows inference on the belief distribution of the agent. To ease notation, in the remainder we denote the mass inside the interval by $M = \Pr(X \in [L, U])$, and the width of the interval by $W = U - L$. We indicate values chosen by a agent who maximizes expected utility with *. The expected reward $Eu(S(L, U, x))$ will be denoted for simplicity by $u(S(L, U, X))$.

2.2 Coverage

Under a given scoring rule, M^* will depend on the distribution X and preferences u. Moreover, under any interval scoring rule that rewards the agent for a correct guess, even if he is punished for choosing a greater width, the upper bound of M^* is 1. To see this, consider a degenerate distribution. With such a distribution the agent can cover all the mass in an arbitrarily small interval. The lower bound of the mass covered by the interval will be called the *coverage* of the rule.

Definition 1 *We say that S has* coverage γ *if $M^*(X, S, u) \geq \gamma$ for all X and all concave u.*

It would be desirable that a rule covers exactly γ, so where $M^*(X, S, u) = \gamma$ for all X and all concave u where this is feasible, so where there exists L and U such that $\Pr(X \in [L, U]) = \gamma$. In fact, intervals that contain precisely $p\%$ of the mass, in this case $p = 100 \cdot \gamma$, are known as *$p\%$ credible intervals* or *prediction intervals*, and there has been considerable interest in their elicitation. In particular, the fractile scoring rule (Murphy and Winkler 1974) is a proper interval scoring rules for eliciting credible intervals for risk neutral agents.

However, as we show below in Corrolary 1 there is no proper interval scoring rule that elicits $p\%$ credible intervals when agents are either risk neutral or risk averse. Because our definition of coverage requires $[L*, U*]$ to always contain the credible interval, one can interpret the elicited interval $[L*, U*]$ as a 100% confidence interval for the $\gamma\%$ credible interval. In the future it may be of interest to loosen this notion of coverage and elicit intervals that contain γ with less than 100% confidence.

2.3 Efficiency

The second feature of the rule of interest to the elicitor is the width of the interval. A reasonable objective for the elicitor is to locate the mass captured inside the interval as precisely as possible, that is, in an interval that is as small as possible. In this context the following notion of efficiency is natural.

Definition 2 *A scoring rule S with coverage γ is called* efficient *within $\mathcal{S}$ if there is no scoring rule $S' \in \mathcal{S}$ with coverage γ such that $W^*(X, S', u) \leq W^*(X, S, u)$ for all $X \in \Delta$ and all concave u with "$<$" for some u and X.*

Definition 2 imposes an intuitive criterion of efficiency saying that S should be *undominated* in terms of the interval width. That is, no other rule should dominate S in that it covers mass γ in a weakly smaller interval for all belief distributions, and in a strictly smaller interval for at least one distribution. If such a rule does not exist, then in this sense S locates mass as precisely as possible.

3 The Truncated Interval Scoring Rule

Having set the stage for evaluating and comparing interval scoring rules, we now propose a specific interval scoring rule S_γ that we call the *truncated interval scoring rule* (TISR). We will show that under some assumptions on X, this rule has the attractive property that it satisfies the efficiency criterion developed above. That is, we can specify its coverage and show that there is no rule with the same coverage that dominates S_γ.

3.1 Definition

We call the scoring rule S_γ the truncated interval scoring rule with parameter $\gamma \in (0, 1)$ if

$$S_\gamma(L, U, x) = \begin{cases} ((b-a) - (U-L))^{\frac{1-\gamma}{\gamma}} & \text{if } x \in [L, U] \text{ and } U - L \leq \gamma \cdot (b-a) \\ 0 & \text{otherwise.} \end{cases}$$

This rule rewards the agent if her guess is correct, i.e. if the realization of the random variable belongs to the specified interval. The parameter γ determines the decrease in the reward as a result of specifying a large interval: a lower γ corresponds to a higher penalty for widening the interval. In section 3.4 we explain the argument behind the truncation, i.e. the fact that the reward drops to zero if the interval is bigger than $\gamma \cdot (b-a)$. Note that S_γ is linear when $\gamma = 1/2$, this particular representative without the truncation has been used in the literature (e.g. Galbiati *et al.* 2009) albeit without deriving its properties. Other papers have used more complicated linear rules with extra penalty parameters, that for example punish the distance of x to the midpoint of the interval (Schmalensee 1976) or to the bounds of the interval (Hamill and Wilks 1995).

3.2 Existence of the optimal interval

We first show that there exists an interval that maximizes the expected utility of the agent under TISR.

Proposition 1 *There exist L^* and U^* with $a \leq L^* \leq U^* \leq b$ such that $u(S_\gamma(L^*, U^*, X)) = \sup_{L,U:a\leq L\leq U\leq b} u(S_\gamma(L, U, X))$.*

The result is obtained by showing that $u(S_\gamma(L, U, X))$ is upper semi-continuous. Then, by the extreme value theorem, it attains a maximum on the compact domain.

3.3 Coverage of the TISR

We now proceed to characterize the coverage of TISR under the following assumption:

$$X \text{ is single-peaked,} \tag{A2}$$

where single peakedness is defined as follows.

Definition 3 *X is single-peaked if there exists $x_0 \in [a, b]$, called a mode of X, such that for any $\varepsilon \geq 0$ we have that $\Pr(X \in [x, x+\varepsilon])$ is increasing in x for $x + \varepsilon \leq x_0$ and decreasing in x for $x \geq x_0$.*

Single-peaked beliefs represent a class of simple belief distributions that are easy to work with and intuitively attractive. Moreover, as we will show below, it allows some crucial simplifications that allow us to derive our results. Note that if X is single-peaked and $\Pr(X = x') > 0$ then $x' = x_0$. Single-peakedness thus implies that X can have at most one point mass and this must be at the mode x_0. Moreover, the density of X must be increasing to the left and decreasing to the right of x_0. An agent with single-peaked beliefs who is paid using TISR will for given width W aim to maximize the mass contained in $[L, U]$. In particular this means that $x_0 \in [L^*, U^*]$.

We will now establish a connection between the parameter γ and the mass inside the interval M^*, using the first order conditions. A crucial variable in this relation is the density on the boundary of the interval. Because the density function need not be differentiable everywhere (although we show that due to single-peakedness it is differentiable almost everywhere), the value of the density on the 'inside' of the boundary may differ from that on the 'outside' of the interval boundary. So notation will be helpful. We denote by $f(y)_- = \lim_{x \uparrow y} f(x)$ and $f(y)_+ = \lim_{x \downarrow y} f(x)$ the respective densities to the left and to the right of the point y. Let $S_\gamma(L, U, in) = S(L, U, x : x \in [L, U])$ be the score when the agent's guess is correct, i.e. when $x \in [L, U]$. We can now rewrite the relevant first order necessary conditions and obtain

$$\max\left\{f(L^*)_-, f(U^*)_+\right\} \leq -\frac{u'(\cdot) * S'_\gamma(L^*, U^*, in) * M^*}{u(\cdot)} \leq \min\left\{f(L^*)_+, f(U^*)_-\right\}, \tag{1}$$

where $u(\cdot) = u(S_\gamma(L, U, in))$. By single-peakedness the mode x_0 is inside the interval, and hence the left term of (1) is the maximum density outside the interval, and the right term is the minimum density inside the interval. The middle term represents the expected cost of expanding the interval, which consists of the marginal disutility of the lower reward that comes with specifying a higher width, normalized by the reward in case the agent gets it right (the denominator). This expression is positive as S_γ is decreasing in W. The left inequality says that to make sure that the agent does not want to expand the interval, the value of the marginal probability that is captured by the expansion should be low enough compared to the relative cost of expansion. Potentially this condition holds with inequality, since the density need not be continuous at the boundary of the interval. A similar logic explains the right inequality. To ensure that the agent does not want to shrink the interval, the density that would be excluded from the interval by shrinking it should be high enough relative to the gain in the reward whenever $x \in [L^*, U^*]$.

The following step will be referred to below as the *flattening argument*. Any values of L_0 and U_0 that satisfy (1) will continue to do so if we flatten the mass outside the interval, so that it is uniformly distributed on $\Omega \setminus [L_0, U_0]$ with density $(1 - M_0)/((b - a) - W_0)$. This is because by single-peakedness, flattening the mass can only reduce $f(L_0)_-$ and $f(U_0)_+$. Similarly, (1) continues to hold if we flatten the mass *inside* the interval, provided $W_0 > 0$, so that it is uniformly distributed on $[L_0, U_0]$ with density M_0/W_0. Again, single-peakedness guarantees that the right hand inequality in (1) continues to hold. This transformation of mass yields a random variable $\widehat{X}$ that has density $(1 - M_0)/((b - a) - W_0)$ on $[a, L_0) \cup (U_0, b]$, density M_0/W_0 on $[L_0, U_0]$ if $W_0 > 0$, and point mass M_0 at $x = L_0$ if $W_0 = 0$.

Thus, the flattening argument tells us that $[L_0, U_0]$ can be supported as an optimal interval $[L^*, U^*]$ for some X if and only if

$$\begin{aligned}\frac{1 - M_0}{(b - a) - W_0} &\leq -\frac{u'(\cdot) * S'_\gamma(L_0, U_0, in) * M_0}{u(\cdot)} \\ &\leq \frac{M_0}{W_0} \qquad \text{if } W_0 > 0,\end{aligned}$$

where $u(\cdot)$ and $S'_\gamma(L_0, U_0, in)$ are defined as before but now as a function of $\widehat{X}$ instead of X. For a risk-neutral decision maker we can rewrite these conditions as

$$\begin{aligned}\frac{1-M_0}{(b-a)-W_0} \leq \left(\frac{1-\gamma}{\gamma}\right)\frac{M_0}{(b-a)-W_0} \\ \leq \frac{M_0}{W_0} \qquad \text{if } W_0 > 0.\end{aligned} \tag{2}$$

From the left hand side inequality (2) we observe a simple relationship between γ and M^* that holds when the decision maker is risk-neutral. The proof in the appendix extends this finding to the risk averse case.

Proposition 2 *The truncated interval scoring rule S_γ has coverage γ when X is single-peaked.*

It should not be surprising that the coverage increases with γ. A higher γ translates into a lower penalty for widening the interval and therefore to a wider interval and greater coverage. It is also easy to verify that a coverage of exactly γ is attained for the uniform distribution, the associated width is then $\gamma(b-a)$. In this sense, the uniform distribution constitutes a 'worst case' for the agent, as she will cover the least mass. Typically though, $M > \gamma$ even when there is an interval I such that $\Pr(X \in I) = \gamma$.

3.4 Efficiency of the TISR

Having characterized the coverage of the TISR, we proceed to derive conditions under which TISR is efficient according to Definition 2. We call S a *simple scoring rule* if $S(L, U, x) = 0$ if $x \notin [L, U]$ and $S(L, U, x) = S(L, U, x')$ if $x, x' \in [L, U]$.

Proposition 3 *S_γ is efficient within the set of simple scoring rules that have coverage γ for all single-peaked X.*

The intuition behind this result comes from the fact that TISR achieves a coverage of exactly γ for a class of distributions that have one point mass and uniform density otherwise. Any rule that is a candidate to dominate S_γ therefore also has to capture exactly γ, or its induced optimal width will be larger. In the proof we show that any rule that achieves exactly γ for those distributions has to be identical to the TISR when $W \leq \gamma(b-a)$. As remarked above, $\gamma(b-a)$ is the maximum interval width that a risk neutral agent will ever specify, which happens under the worst-case scenario that X is uniform. Larger interval widths will therefore be chosen only by risk averse agents.

Efficiency thus provides the rationale for the *truncation*, i.e. the punishment of intervals W that are larger than $\gamma(b-a)$ with payoff 0. This bound insures that risk averse agents can never include more mass in the interval than would be necessary to capture γ in a worst case scenario.

Since TISR is efficient and sometimes covers more than γ it follows that there is no simple scoring rule that always covers exactly γ for all single peaked X and all concave u. In fact, the proof of Proposition 3 reveals that this statement is true even when u is assumed to be risk neutral.

Corollary 1 *There is no simple scoring rule S such that $M^*(X, S, Id) = \gamma$ for all single-peaked X.*

4 Inference from the TISR

The fact that the TISR is efficient means that it locates the fraction γ of the mass of any single-peaked random variable X as precisely as possible. In this section we discuss some more specific inferences that can be made on the basis of the elicited interval. Note that we maintain the restriction to single-peaked X.

4.1 Location

We consider three different notions of the location of a distribution: the expected value or mean, the mode(s) and the median. We investigate in turn how the interval chosen under the TISR relates to these values.

By a counterexample it is easy to see that the interval under the TISR does not necessarily cover the mean of the distribution.

Example 1. Consider $\varepsilon > 0$ and assume that X is distributed such that $\Pr(X = 0) = 1-\varepsilon$ and $f(x) = \varepsilon$ for $x \in (0, 1]$. Note that this distribution is single-peaked and has expected value $EX = \varepsilon/2$. Suppose that $L = 0$. The first order condition for U is $\varepsilon(1 - U^*) = \left(\frac{1-\gamma}{\gamma}\right)(1 - \varepsilon + U^*\varepsilon)$. It follows that $U^* = max\{0, \gamma - (1 - \gamma)\left(\frac{1-\varepsilon}{\varepsilon}\right)\}$. Thus, if $\gamma + \varepsilon \leq 1$ then $U^* = 0$ and the interval does not include the mean of X. $\square$

We now turn to the mode of the distribution, defined in Definition 3. The following result is given without proof as it follows directly from the first order conditions (1) and the single-peakedness assumption.

Proposition 4 *$[L^*, U^*]$ contains a mode of X.*

Note that the optimal interval will not necessarily cover all modes of X. To see this, note that if X is uniformly distributed on $[a, b]$ then each $x \in [a, b]$ is a mode of X.

The next parameter of interest is the median. Whether the interval always contains the median of the distribution depends on the size of the parameter γ. The interval elicited by TISR need not contain the median when $\gamma < 1/2$ (to see this, consider X that is uniformly distributed). Yet the median will be covered if the interval contains at least 50% of the mass, so it is an immediate implication of Proposition 2.2 that the median will fall inside the interval when $\gamma \geq 1/2$.

Corollary 2 *If $\gamma \geq 1/2$, then $[L^*, U^*]$ contains the median.*

Note that this result relies on the use of flattening argument, and thus on the assumption of single-peakedness. One can easily construct examples of non-single-peaked distributions where $\gamma \geq 1/2$ is not sufficient to cover the median.

4.2 Inference on the Dispersion of Beliefs

The width of the interval for a given scoring rule depends on u and X. Regarding X, one would think that an agent will specify a smaller interval, the less dispersed or noisy are her beliefs and the more 'certain' she is about X. Here we establish a formal statement for this intuition. We show how the interval $[L^*, U^*]$ changes when beliefs become more noisy in the following sense.

Definition 4 *X_ε is noisier than X if*

$$X_\varepsilon = \begin{cases} X & \text{with probability } 1-\varepsilon \\ Y & \text{with probability } \varepsilon, \end{cases}$$

where $\varepsilon \in [0,1]$ and Y is uniformly distributed on Ω.

Note that under this notion of noisiness, unlike a mean preserving spread, the expected value typically changes when noise increases.

Proposition 5 *Assume $\gamma \geq 1/2$. If X' is noisier than X, then $W^*(X, S_\gamma, u) \leq W^*(X', S_\gamma, u)$.*

Proposition 5 establishes that a policy maker can use the TISR to get insights into the degree of noisiness or dispersion of the beliefs of the agent. The condition on γ insures that the rule is concave, and the first-order conditions are sufficient. However, unless the preferences of the agent are known, inference about the noisiness of the distribution will be confounded with inferences about the risk aversion of the agent as the next result shows.

We now turn to the effect of u on the interval width, expecting that agents that are less risk averse to specify larger intervals. This intuition can be formalized as follows. We say that $\hat{u}$ is more risk averse than u if there is a concave function g such that $\hat{u}(x) = g(u(x))$ for all x.

Proposition 6 *Assume $\gamma \geq 1/2$. If $\hat{u}$ is more risk averse than u then $W^*(X, S_\gamma, u) \leq W^*(X, S_\gamma, \hat{u})$.*

Proposition 6 tells us that a more risk averse agent will always specify a larger width (and in fact also generate a higher coverage).[1] This is intuitive, since specifying a larger width decreases the probability of getting a payoff of zero. In particular this means that any coverage attained for a risk neutral agent is also attained for a risk averse agent.

In sum, learning about the dispersion of beliefs is confounded. When u can be reasonably held constant, for example by eliciting intervals over time for the same agent, the elicitor can falsify the hypothesis that the beliefs of an agent become noisier. This is important, since the noisiness of the distribution can be interpreted as a proxy of uncertainty, which in many applications will be relevant to the elicitor. In the same vein, if X can be assumed to be constant over agents, for example because they have received the same information, the interval width gives information about their relative degrees of risk aversion.

5 Conclusion

In this article, we have formulated the notions of coverage and efficiency for interval scoring rules. We presented a simple rule, the truncated interval scoring rule, or TISR, that satisfies these criteria under the assumption of single peaked beliefs. The elicitor guarantees that the elicited interval covers at least a mass of γ by choosing the representative that is indexed by this parameter. Moreover, TISR is efficient, in the sense that no other rule achieves a coverage of γ within a weakly smaller interval (and strictly smaller for some distributions). Thus, TISR offers the elicitor the prospect of locating a proportion γ of the mass within the support, and does so in the most precise way possible.

In contrast to other scoring rules, the inferences from the TISR are valid for any degree of risk aversion. The TISR is easy to understand, and requires only two inputs by the agent. This eases constraints on time and cognitive resources of the subjects in economic experiments.

On the basis of these results we believe that interval scoring rules, and the TISR in particular, can be a useful addition to the arsenal of belief elicitation methods. However, to date their properties are not fully understood. A continued investigation of the properties of interval elicitation methods is therefore of interest. An obvious topic for future research is to devise an interval scoring rule for covering the mean. This may be possible if one designs more complicated

[1] The proof of Proposition 6 reveals that

$$[L^*(X, S_\gamma, u), U^*(X, S_\gamma, u)] \subseteq [L^*(X, S_\gamma, \hat{u}), U^*(X, S_\gamma, \hat{u})].$$

rules that include extra penalty parameters. For example, one could punish the distance of x from the midpoint of the interval (Schmalensee 1976) or the bounds of the interval (Hamill and Wilks, 1995).

As we remarked above, the notion of coverage that we propose requires a rule to elicit intervals that cover $\gamma\%$ of the mass for sure. One could loosen this requirement and design rules that cover a mass of γ with less than 100% confidence. Analogously, the notion of eliciting confidence intervals could be extended to proper scoring rules. Known proper scoring rules (like the QSR) elicit point estimates of the parameters of interest, but as we have emphasized, these methods do not apply to general risk preferences. Further research could relax the requirement that confidence intervals have width zero and instead try to elicit a 95% confidence interval for the mean for general risk preferences.

References

Aitchison, J. and I.R. Dunsmore. 1968. "Linear-Loss Interval Estimation of Location and Scale Parameters." *Biometrika*, 55:1, pp. 141-148.

Brier, G.W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, 78, pp. 1-3.

Croson, R. 2000. "Thinking like a game theorist: Factors affecting the frequency of equilibrium play." *Journal of Economic Behavior and Organization*, 41:3, pp. 299-314.

Dufwenberg, M. and U. Gneezy. 2000. "Measuring Beliefs in an Experimental Los Wallet Game." *Games and Economic Behavior*, 30, 163-89.

Fountain, John. 2002. "Eliciting beliefs from risk averse forecasters using a log scoring rule." Working Paper, University of Canterbury.

Galbiati, Roberto, Karl H. Schlag, and Joël J. van der Weele. 2009. " Can Sanctions Induce Pessimism? An Experiment." Labsi Working Paper 24/2009, Siena.

Gordon, Russell A. 1994. *The Integrals of Lebesgue, Denjoy, Perron and Henstock.* AMS Bookstores.

Hamill, T. M. and D. S. Wilks. 1995. "A Probabilistic Forecast Contest and the Difficulty in Assessing Short-Range Forecast Uncertainty." *Weather and Forecasting*, 10, 620-631.

Huck, Stephen and Georg Weizsäcker. 2002. "Do Players Correctly Estimate What Others Do? Evidence of Conservatism in Beliefs." *Journal of Economic Behavior and Organization*, 47, 71-85.

Murphy, Allan, H. and Robert L. Winkler. 1974. "Credible Interval Forecasting: Some Experimental Results." *Monthly Weather Review*, 102:11, 784-94.

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

Rutström, Elizabeth E. and Nathaniel Wilcox. 2008. "Stated versus inferred beliefs: A methodological inquiry and experimental test." MPRA paper No. 11852, Munich.

Schmalensee, Richard. 1976. "An Experimental Study of Expectation Formation." *Econometrica*, 44, pp. 17-41.

Selten, Reinhard, Abdolkarim Sadrieh and Klaus Abbink. 1999. "Money Does Not Induce Risk Neutral Behavior, But Binary Lotteries Do Even Worse". *Theory and Decision*, 46, 211-249.

Winkler, Robert L. 1972. "A Decision-Theoretic Aproach to Interval Estimation." *Journal of the American Statistical Association*, 67:337, pp. 187-91.

Winkler, Robert L. and Allan H. Murphy. 1970. "Nonlinear Utility and the Probability Score." *Journal of Applied Meteorology*, 9, pp. 143-8.

Appendix with Mathematical Proofs

Proof of Proposition 1. By an extension of the extreme value theorem, we know that an upper semi-continuous function attains a maximum on a compact domain. Hence, the proof is complete once we show that $u(S_\gamma(L, U, X))$ is upper semi-continuous in L and U. Note that $u\left(((b-a)-(U-L))^{\frac{1-\gamma}{\gamma}}\right)$ is continuous in L and U. So all we have to show is that $\Pr(X \in [L, U])$ is upper semi-continuous, i.e. for every L_0, U_0 with $L_0 \leq U_0$ and every $\varepsilon > 0$ we need to show that there exists $\delta > 0$ such that $\|(L, U) - (L_0, U_0)\| < \delta$ implies $\Pr(X \in [L, U]) \leq \Pr(X \in [L_0, U_0]) + \varepsilon$. Since $\Pr(X \in [L, U]) \leq \Pr(X \in [\min\{L, L_0\}, \max\{U, U_0\}])$ it is sufficient to prove the claim for $[L, U]$ such that $[L_0, U_0] \subseteq [L, U]$.

Note that $\Pr(X \in [L, U]) = \Pr(X \leq U) - \Pr(X < L)$. Let $F(x) = P(X \leq x)$ be the cdf of X, which is right-continuous and non-decreasing. This implies that $\Pr(X \leq U) = F(U)$ is right continuous in U. Thus, for every $\varepsilon > 0$ there exists $\delta > 0$ such that $U \leq U_0 + \delta$ implies that $\Pr(X \leq U) \leq \Pr(X \leq U_0) + \varepsilon/2$. Similarly, let $F_{-}(x) = P(X < x)$, which is left-continuous and non-increasing. This implies that $\Pr(X < L) = F_{-}(L)$ is left continuous in L. Again, for every $\varepsilon > 0$ there exists $\delta > 0$ such that $L \geq L_0 - \delta$ implies that $\Pr(X < L) \geq \Pr(X < L_0) - \varepsilon/2$. This implies $\Pr(X \in [L, U]) \leq \Pr(X \in [L_0, U_0]) + \varepsilon$, which means that $u(S_\gamma(L, U, X))$ is upper semi-continuous. ■

Proof of Proposition 2. The outline of the proof is as follows. In step 1 we derive some properties of the distribution function of X. In step 2 we derive necessary and sufficient conditions for optimality. In step 3 we reformulate these in conditions that are easier to check. In step 4 we use these to prove the statement of the proposition.

Step 1. Let $F(x) = \Pr(X \leq x)$, so F is monotonically increasing and hence differentiable almost everywhere (see e.g. Gordon 1994, p. 514).[2] Let f be its derivative when it exists, so $f \geq 0$. Since X is single-peaked, there exists x_0 such that f is monotone increasing for $x < x_0$ and monotone decreasing for $x > x_0$. Moreover, X can have at most one mass point, we may assume that this attained at x_0. Let $\xi = \Pr(X = x_0)$. Together, this implies that $F(x) = \int_a^x f(x)\,dx + \xi * \mathbf{I}_{\{x \geq x_0\}}$. Since f is monotone on either side of x_0, it follows that f is differentiable almost everywhere, in particular f is continuous almost everywhere.

Step 2. If F is differentiable at U, $\Pr(X = U) = 0$ and $U - L < \gamma(b-a)$ then $\frac{\partial}{\partial U} u(S_\gamma(L^*, U, X))$ exists and is given by:

$$\begin{aligned} \frac{\partial}{\partial U} u(S_\gamma(L^*, U, in)) &= \frac{\partial}{\partial U}\left(u(S_\gamma(L^*, U, in)) * \Pr(X \in [L^*, U])\right) \\ &= f(U) * u(\cdot) - u'(\cdot) * S'_\gamma(L^*, U^*, in) * \Pr(X \in [L^*, U]) \end{aligned}$$

where

$$S_\gamma(L^*, U^*, in) = ((b-a) - (U-L))^{\frac{1-\gamma}{\gamma}},$$

i.e. the payoff of the agent if $x \in [L^*, U^*]$.

Since F need not be continuous everywhere we formulate the first order necessary conditions in terms of inequalities:

$$\lim_{U \downarrow U^*} \frac{\partial}{\partial U} u(S_\gamma(L^*, U^*, in)) \leq 0 \text{ if } U^* < b, \tag{3}$$

$$\lim_{U \uparrow U^*} \frac{\partial}{\partial U} u(S_\gamma(L^*, U^*, in)) \geq 0 \text{ if } U^* > a. \tag{4}$$

Note that these limits exist as f is monotone on either side of x_0.

Similarly, when f is continuous at L and $\Pr(X = L) = 0$ we can derive

$$\frac{\partial}{\partial L} u(S_\gamma(L, U^*, in)) = -f(U) * u(\cdot) + u'(\cdot) * S'_\gamma(L^*, U^*, in) * \Pr(X \in [L^*, U]),$$

[2] 'Almost everywhere' means that the set of points where F is not differentiable has Lebesgue measure 0.

and therefore

$$\lim_{L \downarrow L^*} \frac{\partial}{\partial L} u\left(S_\gamma\left(L^*, U^*, in\right)\right) \leq 0 \text{ if } L^* < b, \tag{5}$$

$$\lim_{L \uparrow L^*} \frac{\partial}{\partial L} u\left(S_\gamma\left(L^*, U^*, in\right)\right) \geq 0 \text{ if } L^* > a. \tag{6}$$

To simplify notation in the remainder, assume that $f(x) = 0$ if $x \notin [a, b]$ and let $f(x^*)_- := \lim_{x \uparrow x^*} f(x)$ and $f(x^*)_+ := \lim_{x \downarrow x^*} f(x)$. Recall that we defined $M^* = \Pr\left(X \in [L^*, U^*]\right)$ and $W^* = U^* - L^*$. It is easy to see that $u\left(S\left(L^*, U^*, in\right)\right) > 0$. It follows then from (3), (4), (5) and (6), that

$$\max\left\{f\left(L^*\right)_-, f\left(U^*\right)_+\right\} \leq \frac{u'(\cdot)}{u(\cdot)} S'_\gamma(L^*, U^*, in) * M^* \leq \min\left\{f\left(L^*\right)_+, f\left(U^*\right)_-\right\}. \tag{7}$$

Step 3. Since $S_\gamma\left(W^*\right) = \left((b-a) - W^*\right)^{\frac{1-\gamma}{\gamma}} > 0$ it follows that $\left(L^*, U^*\right) \neq (a, b)$. So we can assume without loss of generality that $L^* > a$. Since f is increasing on $[a, L^*)$ and decreasing on $(U^*, b]$ we obtain that $f(x) \leq \max\left\{f\left(L^*\right)_-, f\left(U^*\right)_+\right\}$ for all $x \in [a, L^*) \cup (U^*, b]$. Therefore (7) continues to hold if we "flatten" the mass outside $[L^*, U^*]$ to obtain that $f(x) = q_o^*$ for $x \in [a, L^*) \cup (U^*, b]$ where

$$q_o^* = \frac{Pr\left(X \notin [L^*, U^*]\right)}{(b-a) - (L^* - U^*)} = \frac{1 - M^*}{(b-a) - W^*}. \tag{8}$$

Similar arguments show that (7) continues to hold if we "flatten" mass inside of $[L^*, U^*]$ to obtain $f(x) = q_i^*$ for $x \in [L^*, U^*]$ where $q_i^* = M^*/W^*$. (7) can now be rewritten as

$$q_o^* = \frac{1 - M^*}{(b-a) - W^*} \leq \frac{u'(\cdot)}{u(\cdot)} S'_\gamma(L^*, U^*, in) * M^* \leq \frac{M^*}{W^*} = q_i^*. \tag{9}$$

If $u(x) = x$ then these conditions reduce to

$$\frac{1 - M^*}{(b-a) - W^*} \leq \frac{1-\gamma}{\gamma} \frac{1}{(b-a) - W^*} M^* \leq \frac{M^*}{W^*}. \tag{10}$$

Step 4. Since u is concave, $u'(z)/u(z) \leq 1/z$ for $z > 0$ and hence

$$\frac{u'(\cdot)}{u(\cdot)} S'_\gamma(L^*, U^*, in) * M^* \leq \frac{1}{(b-a) - W^*} \left(\frac{1-\gamma}{\gamma}\right) M^*. \tag{11}$$

Combining (11) with the left hand side in (9) yields

$$\frac{1 - M^*}{(b-a) - W^*} \leq \frac{1}{(b-a) - W^*} \left(\frac{1-\gamma}{\gamma}\right) M^*$$

and hence

$$M^* \geq \gamma.$$

■

Proof of Proposition 3. Assume that S_γ is not efficient among the simple scoring rules. So there exists a simple scoring rule S with coverage γ that satisfies $W^*(X, S, u) \leq W^*(X, S_\gamma, u)$ for all $X \in \Delta$ and all concave u with strict inequality for some X and some u. In the following we will show that S is identical to S_γ up to a constant factor when $U - L \leq \gamma(b-a)$. Since $W^*(X, S_\gamma, u) \leq \gamma(b-a)$ this then implies that $W^*(X, S, u) = W^*(X, S_\gamma, u)$ for all X and u which contradicts the above and hence proves that S_γ is efficient among the simple scoring rules.

To simplify exposition we will assume that $a = 0$ and $b = 1$. Note that this can be done without loss of generality by appropriate rescaling of the scoring rule.[3] Consider the class of random variables X_z indexed by z for $z \in [0, \gamma]$ where the underlying distribution F_z puts point mass $\frac{\gamma - z}{1-z}$ on $x = 0$ and has density $f_z(x) = \frac{1-\gamma}{1-z}$ for $x \in [a, b]$. So $F_z(z) = \gamma$. It follows from (10) that $W^*(X_z, S_\gamma, Id) = z$ so $M^*(X_z, S_\gamma, Id) = \gamma$. As S_γ covers exactly γ of the mass of X_z, so must S, which means that $L^*(X_z, S, Id) = 0$ and $U^*(X_z, S, Id) = z$ and hence $W^*(X_z, S, Id) = z$ for all $z \in [0, \gamma]$. Let $r_z(U) = S(0, U, 0)$. Given $0 < z \leq \gamma$, the first order conditions imply $F_z(z) \cdot r_z'(z) + f_z(z) r_z(z) = 0$ and hence $\gamma r_z'(z) + \frac{1-\gamma}{1-z} r_z(z) = 0$. We solve this first order differential equation and obtain $r_z(z) = c \cdot (1-z)^{(1-\gamma)/\gamma}$ for $z \in (0, \gamma]$ and some $c > 0$. It follows that $S(0, U, in) = c \cdot S_\gamma(0, U, in)$ for all $0 < U \leq \gamma$.

We now show that this result does not only hold for $L^* = 0$, but for more general L^*. Consider now the class of distributions that have point mass $\frac{\gamma - z}{1-z}$ at γ and density $f_z(x) = \frac{1-\gamma}{1-z}$ for $x \in [a, b]$. Due to the uniform mass we can assume that $U^*(X_z, S, Id) = U^*(X_z, S_\gamma, Id) = \gamma$. Replicate the above arguments, defining $r_z(L) = S(L, \gamma, in)$, to show that $S(L, \gamma, in) = c' \cdot S_\gamma(L, \gamma, in)$ for some $c' > 0$. Combining this with our previous analysis, setting $L = 0$, shows that $c = c'$.

Continuing this way one can show, tediously, that $S(L, U, in) = c \cdot S_\gamma(L, U, in)$ whenever $U - L \leq \gamma$ which completes the proof. ■

Proof of Proposition 5. Consider random variables X, Y and X_ε as in Definition 4. Let $[L_0^*, U_0^*]$ be the optimal interval selected under X and $W_0^* = U_0^* - L_0^*$. We know from the first order conditions (7) that

$$f(x) \geq \frac{u'(\cdot) * S_\gamma'(L^*, U^*, in)}{u(\cdot)} \Pr(X \in [L_0^*, U_0^*]) \text{ for all } x \in [L_0^*, U_0^*]. \tag{12}$$

We want to show that this also holds for X_ε at W_0^*, hence that

$$f_\varepsilon(x) \geq \frac{u'(\cdot) * S_\gamma'(L^*, U^*, in)}{u(\cdot)} \Pr(X_\varepsilon \in [L_0^*, U_0^*]) \text{ for all } x \in [L_0^*, U_0^*].$$

[3] If S is a scoring rule for $X \in \Delta[0,1]$ with coverage γ then S' is a scoring rule for $X \in \Delta[a,b]$ with the same coverage if $S'(L, U, x) = S\left(\frac{L-a}{b-a}, \frac{U-a}{b-a}, \frac{x-a}{b-a}\right)$.

Since $\gamma \geq 1/2$ and hence the first order condtions are sufficient, this then implies that $[L_0^*, U_0^*]$ is contained in $[L^*(X_\varepsilon, S_\gamma, u), U^*(X_\varepsilon, S_\gamma, u)]$.

We can write

$$\Pr(X_\varepsilon \in [L_0^*, U_0^*]) = (1-\varepsilon)\Pr(X \in [L_0^*, U_0^*]) + \frac{\varepsilon W_0^*}{b-a}, \tag{13}$$

$$f_\varepsilon(x) = (1-\varepsilon) f(x) + \frac{\varepsilon}{b-a}\ . \tag{14}$$

Using (13) and (14), we have (with some suppression of notation)

$$\begin{aligned} f_\varepsilon(x) - \frac{u'(\cdot) * S_\gamma'(L^*, U^*, in)}{u(\cdot)} \Pr(X_\varepsilon \in [L_0^*, U_0^*]) \\ = (1-\varepsilon)\left[f(x) - \frac{u'(\cdot) * S_\gamma'(L^*, U^*, in)}{u(\cdot)} \Pr(X \in [L_0^*, U_0^*])\right] \\ + \varepsilon\left[\frac{1}{b-a} - \frac{u'(\cdot) \cdot S_\gamma'(L^*, U^*, in)}{u(\cdot)} \frac{W_0^*}{(b-a)}\right] \end{aligned}$$

We want to show that this expression is positive. We know from (12) that the first term between square brackets is positive. Remains to show that

$$\frac{u'(\cdot) * S_\gamma'(L^*, U^*, in)}{u(\cdot)} \frac{W_0^*}{(b-a)} \leq \frac{1}{b-a}\ . \tag{15}$$

But this is equivalent to the right hand side of (9). ■

Proof of Proposition 6. Again we use the first order conditions which, given $\gamma \geq 1/2$, are sufficient. Consider concave functions u, $\hat{u}$ and g such that $\hat{u}(x) = g(u(x))$. Let $W_0^* = W^*(X, S_\gamma, u)$. Following (7),

$$f(x) \geq \frac{u'(\cdot)}{u(\cdot)} S_\gamma'(L^*, U^*, in) \cdot \Pr(X \in [L_0^*, U_0^*]) \text{ for all } x \in [L_0^*, U_0^*].$$

In order to prove $W^*(X, S_\gamma, u) \leq W^*(X, S_\gamma, \hat{u})$ it is enough to show that

$$f(x) \geq \frac{\hat{u}'(\cdot)}{\hat{u}(\cdot)} S_\gamma'(L^*, U^*, in) \cdot \Pr(X \in [L_0^*, U_0^*]) \text{ for all } x \in [L_0^*, U_0^*].$$

Since g is concave, $g'(x)/g(x) \leq 1/x$, and hence

$$\frac{\hat{u}'(x)}{\hat{u}(x)} = \frac{g'(u(x))\, u'(x)}{g(u(x))} \leq \frac{u'(x)}{u(x)}$$

which implies the desired inequality. ■