Incentives for eliciting confidence intervals

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

DOI
10.2139/ssrn.2271061
Publication date
2012
Document Version
Final published version

Incentives for Eliciting Confidence Intervals

Karl H. Schlag* Joël J. van der Weele†

December 3, 2012

Abstract

A natural way to obtain information about the concentration and dispersion of an expert's beliefs is to ask for a confidence interval. Our objective is to design an elicitation mechanism that rewards the expert on the basis of the realized event and satisfies a set of desirable properties. We show that the existing mechanisms fail some of these properties, and formulate a new mechanism - the Truncated Interval Scoring Rule - that has all properties and is easily implementable in experimental work.

Keywords: Belief elicitation, scoring rules, subjective probabilities, confidence intervals.
JEL Codes: C60, C91, D81.

*University Vienna, Vienna. E-mail: karl.schlag@univie.ac.at
†Corresponding author. J.W. Goethe University, Grüneburgplatz 1, RuW Gebäude 4. Stock, 60323 Frankfurt am Main, Germany. Tel. +49 (0)69 79834814. E-mail: vanderweele@econ.uni-frankfurt.com

1 Introduction

We consider belief elicitation mechanisms where the expert specifies a confidence interval and is paid according to a realization of the event. Confidence intervals are a natural way to get insight into the events an expert thinks likely to occur and the dispersion of the expert's beliefs. Moreover, interval elicitation is relatively simple, since it only requires the elicitation of two numbers. It is a common practice in climate predictions and weather reporting, forecasting of financial variables like inflation or asset prices and is gaining popularity in economic experiments (e.g. Schmalensee, 1976; Kirchler and Maciejovsky, 2002; Bottazzi, Devetag, and Pancotto, 2011).

From the standpoint of the elicitor, it is desirable to provide the expert with rewards for specifying a truthful interval. Such incentives make experts take the forecast seriously and perhaps gather additional information. Incentives can also help the expert to think in a systematic way about the trade-offs that are important to the elicitor and help align the objectives of the elicitor and the expert. This idea has been influential in the literature on the elicitation of subjective probabilities and means, and has produced a large theoretical and experimental literature on 'scoring rules' (Offerman, Sonnemans, Van de Kuilen, and Wakker, 2009; Gneiting and Raftery, 2007).[1] By contrast, the literature on interval elicitation is much smaller, and has not generated an accepted experimental methodology.

In this paper we consider the design of reward mechanisms for truthful interval elicitation. We propose four desirable properties of such reward mechanisms and formulate a new scoring rule that has these properties. The first property is accuracy: the expert's optimal interval should cover an amount of probability mass that is predetermined by the elicitor. The second property is that the specified interval contains the events that the expert thinks most likely to occur. We believe that this is a more natural objective for interval elicitation than tracking the mass in the tails of the distribution, which is the approach in the existing literature. The third property is generality: the first two properties should hold for risk averse as well as risk neutral experts. This contrasts with most of the scoring rule literature, which is based on the assumption that the expert is risk neutral. Finally, we require precision: among the rules that satisfy the first three properties, we favor the rule that pins down the mass

[1] Scoring rules have been used for two distinct objectives. First, they can incentivize the expert about some parameter of interest only known to the expert. The second objective, central in the statistics literature, is the ex-post evaluation of the performance of an expert. We focus on the first objective.

in the smallest interval. We evaluate the precision of a mechanism for the 'worst case' beliefs where the interval is the widest.

We show that the interval scoring rules that have been considered in the literature satisfy the first and third objective, but violate the second and the fourth. As a consequence, the expert has an incentive to specify intervals that are too wide, and may exclude the most likely events. We propose a new scoring rule, called the Truncated Interval Scoring Rule (TISR), that satisfies all our four criteria. The rule is simple and confronts the expert with an intuitive trade-off; payment occurs only for a correct prediction, and decreases monotonically in the width of the interval. The TISR thus provides an attractive method for elicitation, and variations based on this paper have already been implemented in economic experiments (Galbiati, Schlag, and Van der Weele, 2011; Cettolin and Riedl, 2011a,b; Riedl and Smeets, 2011; Peeters, Vorsatz, and Walzl, 2012).

Throughout the paper, we only consider an environment where the random variable about which beliefs are elicited has support in a bounded range, as is the case in experimental settings and in many other applications. Moreover, in most of the paper we restrict attention to single-peaked distributions. We believe this is a reasonable assumption in most cases, and it is intuitive that intervals may not be very informative if beliefs are assumed to have multiple peaks. In Section 7 we drop this assumption and select a rule that elicits sets instead of a single interval.

To our knowledge, there has been no systematic effort to compare and select incentives for interval elicitation for a general class of belief distributions. Schmalensee (1976) and Winkler and Murphy (1979) provide scoring rules that we discuss in Section 4. Aitchison and Dunsmore (1968) and Winkler (1972) consider optimal intervals under more general piece-wise linear scoring rules, where Aitchison and Dunsmore (1968) assume that the scale parameter (variance) of the underlying distribution is known. There is a statistical literature that uses the first and fourth criterion explained above within a smaller class of belief distributions, such as normal distributions (Casella and Hwang, 1991).

The influence of this theoretical literature on experimental practice in economics and psychology has been limited. Most researchers use unincentivized elicitation, and some have applied reward mechanisms that do not provide incentives for truthful belief reporting.[2] This is unfortunate, since there is evidence that experimental subjects

[2]For example, Cesarini, Sandewall, and Johannesson (2006) reward the subjects if they correctly estimate the hit rate of their previously stated intervals. Blavatskyy (2008) shows that this method

often do not share the elicitation objectives of the elicitor, and that this problem can be ameliorated with incentives for interval elicitation. Countless studies show that 90% confidence intervals elicited without incentives are accurate much less than 90% of the time (Moore and Healy, 2008). Yaniv and Foster (1995, 1997) show that this occurs because subjects purposefully specify excessively narrow intervals in order to provide more informative forecasts. Krawczyk (2011) experimentally compares interval elicitation without and with incentives, using the scoring rule of Winkler and Murphy (1979). Accuracy in the incentivized condition is significantly closer to the confidence level set by the elicitor.

The paper proceeds as follows. In the next section we outline the environment under consideration. In Section 3 we propose desirable properties of interval scoring rules. In Section 4 we evaluate the properties of existing rules in the literature. In the second part of the paper we propose a simple scoring rule and evaluate its performance. In the conclusion we summarize the performance of the new and the existing rules and discuss further research. All proofs are in the Appendix.

2 Preliminaries

We consider an expert endowed with preferences over $\mathbb{R}$ that admit an expected utility representation, denoted by u, that is contained in a given set of possible utility functions denoted by $\mathcal{U}$. This expert has subjective beliefs over the distribution of a random variable X with realization x and cdf F_X. The distribution F_X is only known to the expert, but the elicitor knows that F_X generates outcomes belonging to $[a, b]$.

We assume that $F_X \in \Delta$, where we assume Δ to be the class of all single-peaked distributions. In Section 7 we will consider more general distributions and the elicitation of a set instead of an interval. Single-peakedness is defined as follows.

Definition 1 *F_X is single-peaked if there exists $x_0 \in [a, b]$ 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-peakedness implies that F_X can have at most one mass point. Its density function will be increasing when $x < x_0$ and decreasing when $x > x_0$, where x_0 is also called a mode of F_X.

is easy to game. Other studies (e.g. Budescu and Du (2007)) simply reward subjects proportional to their accuracy rate, which can be gamed by simply reporting very large intervals regardless of beliefs.

We consider an elicitor who asks an expert to specify boundaries L and U of an interval, where $L \leq U$.[3] The elicitor then pays the expert an amount $S(L, U, x)$ after drawing a realization x from the random variable of interest. This payment function S is also called a *scoring rule*. Formally, a scoring is a function $S : [a, b]^3 \to \mathbb{R}$. The set of all such scoring rules is denoted by $\mathcal{S}$. We will let $u(S(L, U, X))$ denote the expected utility when specifying L and U given X, so $u(S(L, U, X)) = \int_a^b u(S(L, U, x)) \, dF_X(x)$. Let $W = U - L$ be the width of the interval, and $M = \Pr(X \in [L, U])$ the mass covered by the interval. We denote values that maximize the expected utility of the expert by *, and write these as functions of F_X, S and u whenever necessary.

3 Properties for Interval Elicitation Mechanisms

We propose four desirable properties for interval scoring rules. This list includes some known properties such as coverage and minmax width, and defines new ones. While we think that these four properties are fundamental to interval elicitation, we do not claim that this list is exhaustive. For example, experimentalists may be interested in aspects of the scoring rule that relate to implementability, such as a bounded score and simplicity. Although simplicity is a somewhat subjective property, we will discuss it in more an less formal ways throughout the paper.

3.1 Accuracy

The first property is that the elicited interval is a $\gamma{\cdot}100\%$ confidence interval, so that the elicitor knows how much mass is being captured. This property can be formalized in several ways, the first one being well-known in the literature on scoring rules.

Definition 2 (Proper) *S is "proper" if $M^*(F_X, S, u) = \gamma$ for all $F_X \in \Delta$ and all $u \in \mathcal{U}$.*

Naturally, properness is a valuable characteristic of a rule, since the elicitor knows exactly how much mass is within the interval. For the case of eliciting probabilities

[3] Our discussion does not include the elicitation of entire (continuous) belief distributions, from which intervals can be calculated as a byproduct (Winkler and Matheson, 1976). Such rules require the subject to state an entire probability density function, which is time-consuming, and is little used in experimental economics. By contrast, interval elicitation only requires the subject to report two numbers. Furthermore, the attractiveness of these rules relies on the assumption of risk neutrality, an assumption we relax in this paper.

and means, Schlag and van der Weele (2012) shows that there is no proper rule if $\mathcal{U}$ contains all risk averse as well as risk neutral preferences, and we conjecture a similar result to hold for interval elicitation. Therefore, following Casella and Hwang (1991), we formulate a weaker property.

Definition 3 (Coverage) *S has "coverage γ" if $M^*(F_X, S, u) \geq \gamma$ for all $F_X \in \Delta$ and all $u \in \mathcal{U}$.*

Note that this definition of coverage, like the definition of the level of a statistical test, implies that a rule with coverage γ also has coverage γ', for all $\gamma' \leq \gamma$.

3.2 Most Likely

If the goal of belief elicitation is to understand what the expert thinks will happen, the stated interval should not only include a mode of the belief distribution, but the events it includes should be at least as likely to occur as events that it does not include. This is captured by our 'most likely' property:

Definition 4 (Most likely) *S elicits "most likely events" if for all $F_X \in \Delta$ and $u \in \mathcal{U}$ there is no $\varepsilon > 0$ and intervals $[a_1, a_1 + \varepsilon]$ and $[a_2, a_2 + \varepsilon]$ such that $[a_1, a_1 + \varepsilon] \subseteq [L^*, U^*]$ and $[a_2, a_2 + \varepsilon] \cap [L^*, U^*] = \emptyset$ and $Pr(X \in [a_1, a_1 + \varepsilon]) < Pr(X \in [a_2, a_2 + \varepsilon])$.*

This property will be less relevant if the goal of elicitation is not to find out what the expert thinks will occur, but instead to obtain information about tail risks.

3.3 Generality

Faced with a scoring rule, the expert's optimal interval depends on her beliefs and risk preferences. Inferences from interval elicitation that depend on strong assumptions on the risk preferences of the expert should be approached with caution. Specifically, scoring rules for the elicitation of probabilities typically rely on the assumption of risk neutrality, which we regard as empirically implausible. Holt and Laury (2002) presents evidence that most experimental subjects are risk averse. Armantier and Treich (2010) and Offerman, Sonnemans, Van de Kuilen, and Wakker (2009) show that most subjects behave as if they are risk averse in the context of belief elicitation.[4]

[4]Several authors have shown how to elicit probabilities without the assumption of risk neutrality, but these methods have their own drawbacks. Offerman, Sonnemans, Van de Kuilen, and Wakker

In light of this evidence, we believe that inference from interval elicitation should be valid for risk averse as well as risk neutral experts.

Assumption 1 *$\mathcal{U}$ is the class of all utility functions that are strictly increasing, continuously differentiable, and concave.*

Continuous differentiability is only assumed for convenience, all proofs in the remainder easily extend to the class of continuous utility functions.

3.4 Precision

Among the rules that satisfy our first three criteria, we would like to pick the one that pins down the mass γ with most *precision*. Precision can be measured by the width of the expert's optimal interval for given beliefs F_X and preferences u. The question is how one should aggregate over all possible beliefs and preferences. A starting point is to discard rules that always induce higher widths than some other rule.

Definition 5 (Admissible) *S with coverage γ is "admissible within $\mathcal{S}$" if there is no scoring rule $\tilde{S} \in \mathcal{S}$ with coverage γ such that*

$$W^*\left(F_X, \tilde{S}, u\right) \leq W^*\left(F_X, S, u\right)$$

for all $F_X \in \Delta$ and all $u \in \mathcal{U}$, with "$<$" for some u and F_X.

Admissibility is a rather weak requirement, since a rule only has to induce a small width for a single combination of beliefs and preferences in order to pass it. A more stringent requirement, used in Casella and Hwang (1991), is to to measure precision in terms of the 'worst case' belief distribution that induces the maximal interval width, and select the rule that minimizes this maximal width.

Definition 6 (Minmax width) *S with coverage γ attains "minmax width within $\mathcal{S}$" if there is no scoring rule $\tilde{S} \in \mathcal{S}$ with coverage γ such that*

$$\sup_{F_X \in \Delta, u \in \mathcal{U}} w\left(F_X, \tilde{S}, u\right) < \sup_{F_X \in \Delta, u \in \mathcal{U}} w\left(F_X, S, u\right).$$

When evaluating precision, we will mostly focus on the property of minmax width.

(2009) proposes a mechanism to correct for each subject's risk aversion individually, but the mechanism is time consuming. Schlag and van der Weele (2012) discusses the literature of inducing risk neutrality by paying in lottery tickets, but it is an open question if these complicated mechanisms work in practice.

4 Applying Criteria to Existing Rules

The literature on scoring rules for belief elicitation focuses on the elicitation of point beliefs rather than intervals. In this section we identify and investigate the properties of the only two scoring rules for interval elicitation that have been justified in terms of desirable properties.[5]

The first rule is due to Winkler and Murphy (1979, WM79 hereafter). It is applied in Hamill and Wilks (1995) and Krawczyk (2011), and discussed in some detail in Gneiting and Raftery (2007). Up to an affine transformation, this rule is given by

$$S_{WM79}(L,U,x) = -\left(\frac{1-\gamma}{2}\right)(U-L) - (L-x)\,1_{\{x<L\}} - (x-U)\,1_{\{x>U\}},$$

where 1_E is an operator that is 1 if the event E is true and 0 otherwise. In words, this rule punishes the expert for specifying a larger interval width, and for the distance of x from the interval bound if x is outside the interval.

The second scoring rule is proposed in Schmalensee (1976, S76 hereafter). Up to an affine transformation, it is given by

$$S_{S76}(L,U,x) = -\left(\frac{1-\gamma}{2}\right)(U-L) - (L-x)\,1_{\{x<L\}} - (x-U)\,1_{\{x>U\}} - \left|x - \frac{L+U}{2}\right|.$$

This rule is similar to S_{WM79}, but it adds an extra penalty if the realization is inside the interval, but away from the mid-point.

4.1 Accuracy

The main reason these rules have been discussed in the literature is that they are proper if the expert is risk neutral. Winkler and Murphy (1979) shows that S_{WM79} elicits the $\frac{1-\gamma}{2}$ and $\frac{1+\gamma}{2}$ quantiles, thus tracking the mass in the tails of the distribution.

Although properness is a desirable property, the assumption that people are risk neutral violates the generality property. Turning to the weaker criterion of coverage, we are able to show a new result.

[5] A third rule suggested by Casella and Hwang (1991) is used with some variations to elicit parameters of normal distributions. It is defined by $S_{CH91}(L,U,x) = 1_{\{L\leq x\leq U\}} - k(U-L)$. This rule does not have good properties in our setting with general distributions. For instance, in order to have coverage when beliefs are uniformly distributed on $[a,b]$ one needs $k>1$, but this implies that $[L,U]=[a,b]$.

Proposition 1 *S_{S76} and S_{WM79} have coverage γ.*

Given the properness of the rule for risk neutral agents, the key to the proof is to show that risk averse experts will always specify a larger interval than risk neutral ones. Note that Schmalensee (1976) proved coverage for S_{S76} for the more restrictive case where the belief distribution is symmetric.

4.2 Most Likely

Neither S_{S76} nor S_{WM79} has the most likely property. To see this, consider a skewed distribution with density $f(x) = \frac{1}{2\sqrt{x}}$, depicted in Figure 1. In the figure, we indicate the optimal intervals for a risk neutral expert under S_{S76} and S_{WM79} (and also of the rule called the TISR that we present in the next section). The example shows that

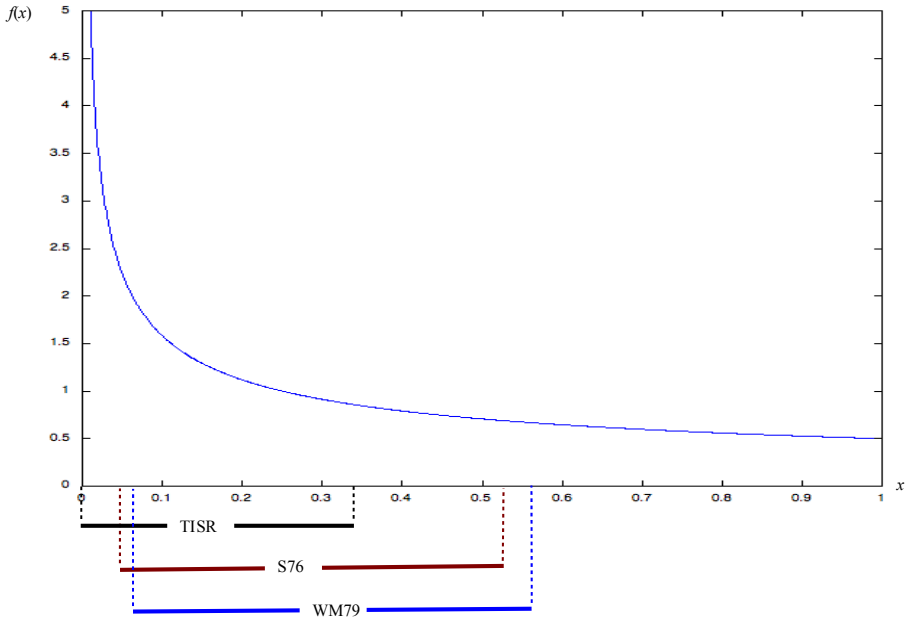

Figure 1: Optimal intervals for TISR, S67, and WM79 when $f(x) = \frac{1}{2\sqrt{x}}$, $\gamma = 0.5$.

under S_{S76} and S_{WM79}, the elicitor cannot infer from the stated interval which events the expert thinks are most likely. The reason is that these rules do not reward the expert for a correct prediction, but 'punish' the expert if the realization is very far from the chosen interval bounds. This means that the expert does not want to specify an interval too far away from either end of the range.

4.3 Precision

The properties of precision and most likely are interdependent. A rule that does not elicit the most likely events will induce a larger interval width to obtain coverage. If this happens for distributions that are close to the 'worst case' distribution, the rule may violate minmax width, as the following result shows.

Proposition 2 *Neither S_{S76} nor S_{WM79} attains minmax width.*

The proof of this result follows from the fact that both S_{S76} and S_{WM79} attain a maximum width of at least $\gamma(b-a)\left(\frac{2}{1+\gamma}\right)$. Below, we show that this is larger than the maximal width of $\gamma(b-a)$ attained by the rule called TISR, specified in the next section. Thus, both S_{S76} and S_{WM79} are imprecise, in the sense that it is in the interest of the elicitor to specify an interval that is too large.

To summarize, the two existing rules are proper for risk neutral experts and have coverage for all concave utility functions and single peaked belief distributions. However, they do not attain minmax width and fail to elicit the most likely events. Note that these two violations occur even if the expert is assumed to be risk neutral.

5 The Truncated Interval Scoring Rule

In this section we propose a new rule that has all four properties specified in Section 3. We will argue that this rule is intuitive and suitable for interval elicitation in experiments.

5.1 Definition and Existence

The Truncated Interval Scoring Rule (TISR) with parameter $\gamma \in (0,1)$ is given by

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

The properties of the rule are invariant to any affine transformation. When $a=0$, $b=1$ and $\gamma=\frac{1}{2}$ the rule is particularly simple:

$$S_T(L,U,x) = \begin{cases} 1-W & \text{if } x \in [L,U] \text{ and } W \leq \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Here, the term 'truncated' refers to the restriction that there is no payment if the expert specifies an interval larger than a fraction γ of the range $[a, b]$. The rationale is that for the worst-case uniform distribution, this fraction covers exactly γ, while for other single-peaked belief distributions one can cover γ in a smaller interval. Thus, TISR punishes the expert for specifying a range that is larger than necessary to obtain coverage.[6]

If the truncation is dropped from the definition we obtain an even simpler scoring rule, called the Interval Scoring Rule (ISR), which has similar properties to TISR. All results in the remainder of this paper are valid for the ISR, except those relating to precision since interval widths may now be larger than necessary to obtain coverage.

The next result shows that the objective of finding an interval that maximizes the score achieved under TISR is well defined.

Proposition 3 *For any $F_X \in \Delta$ there exist L^* and U^* with $a \leq L^* \leq U^* \leq b$ such that $u\left(S_T\left(L^*, U^*, X\right)\right) = \sup_{L,U:a\leq L\leq U\leq b} u\left(S_T\left(L, U, X\right)\right)$.*

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

5.2 Accuracy

With respect to accuracy, we can prove the following.

Proposition 4 *S_T has coverage γ.*

The fact that coverage increases with γ is intuitive, since a higher γ translates into a lower penalty for widening the interval. We give a short sketch of the intuition behind the proof, which is contained in the appendix. Denote by $M(w)$ the maximal subjective probability that can be covered by an interval for a given width w. Then the maximal expected utility of specifying an interval with width w is equal to $u(h(w))M(w)$ where $h\left(w\right) = \left(1 - w\right)^{\frac{1-\gamma}{\gamma}}$. The first order condition related to the optimal choice of the width w is:

$$\frac{d}{dw}\left(u\left(h\left(w\right)\right)M\left(w\right)\right) = M'\left(w\right)u\left(h\left(w\right)\right) - M\left(w\right)u'\left(h\left(w\right)\right)\left(\frac{1-\gamma}{\gamma}\right)\frac{h\left(w\right)}{1-w}. \quad (1)$$

[6]Note that the possibility of truncation relies on the fact that TISR elicits the most likely events. Applying the truncation to S_{S76} and S_{WM79}, which do not have this property, will result in a violation of coverage for some distributions.

The first argument of the RHS of (1) is the marginal benefit of expanding the interval, which consists of an increased likelihood of capturing the realized event. The second term is the marginal cost of doing so, which consists of a decreased payment if the realized event is in the interval. We know that u is concave (by assumption) and M is concave in w because of single-peakedness. Using these facts, we show in the appendix that $M(w) < \gamma$ implies that the derivative with respect to the width (1) is positive, so that the expert would like to expand the interval.

5.3 Most Likely

The next result sets the TISR apart from the existing rules, discussed in the previous section.

Proposition 5 *S_T elicits the most likely events.*

The result is trivial to prove: if the interval does not contain the most likely events, the expert could improve his score by moving the interval. Thus, unlike the existing rules, TISR elicits the most likely events for skewed distributions also. The key to this difference is that the TISR does not punish the size of a failure, so experts have no reason so specify 'cautious' intervals in the middle of the range.

5.4 Precision

TISR attains minimax width among the set of all scoring rules.

Proposition 6 *S_T attains minmax width in $\mathcal{S}$, where*

$$\sup_{F_X \in \Delta, u \in \mathcal{U}} W^* \left(F_X, S_T, u\right) = \gamma(b-a).$$

The proof is simple: In order to cover mass γ under the worst case uniform distribution one needs to have an interval width of at least $\gamma(b-a)$. Hence the maximal width of any rule is at least this number. TISR, by its definition, never elicits a larger interval, and hence attains minmax width.

We can also compare TISR to other rules in terms of admissibility. In Section 7 we show that TISR is not admissible in the set $\mathcal{S}$ of all scoring rules, since there exist more complicated scoring rules that do better. However, we can show admissibility within a smaller set of 'simple' scoring rules. We call S a *simple scoring rule* if there

exists a continuous and decreasing function $h_1 : [0, b-a] \to \mathbb{R}_0^+$ such that $S(L,U,x) = h_1(U-L)$ if $x \in [L,U]$ and $S(L,U,x) = 0$ if $x \notin [L,U]$. So the payoff is positive only if the realization of X lies in the specified interval, and this payoff is a decreasing function of the width of the interval. Let $\mathcal{S}_{simple}$ be the set of such simple scoring rules. Note that TISR is a simple scoring rule.

Proposition 7 *S_T is admissible within $\mathcal{S}_{simple}$.*

In view of the experimental applications of scoring rules, we think a focus on simple rules is desirable and the class of simple rules provides an important benchmark. Specifically, the restriction that payoffs are 0 when x is outside the interval simplifies the exposition of the rule and understandability to experimental subjects.

6 Inference from the TISR

In this section we discuss some specific inferences that can be made from the stated interval when TISR is used.

6.1 Mode, Median and Mean

A primary variable of interest is the mode of the distribution. The following result follows directly from Proposition 5.

Corollary 1 *The interval $[L^*, U^*]$ induced by TISR contains a mode of X.*

This result is not true for WM79 and S76 and we do not know of any scoring rule that elicits the mode of a continuous distribution. Note that the TISR will not necessarily cover all modes of X. For example, if X is uniformly distributed on $[a,b]$ then each $x \in [a,b]$ is a mode of X.

Another parameter of interest is the median. The interval always contains the median if it contains at least 50% of the mass. Therefore, if a scoring rule has coverage γ, then $[L^*, U^*]$ contains the median if $\gamma \geq \frac{1}{2}$.

Corollary 2 *The interval $[L^*, U^*]$ induced by TISR, S_{S76} and S_{WM79} contains the median if $\gamma \geq \frac{1}{2}$.*

Finally, consider the expected value or mean of X. The example below shows that TISR does not cover the mean for sufficiently skewed distributions. For such distributions the mean does not necessarily provide a good indicator of the concentration of mass, so we consider its elicitation an alternative objective to eliciting the most likely events.

Example 1. Consider $\varepsilon > 0$ and assume that X is distributed such that $\Pr(X = 0) = 1 - \varepsilon$ and $f_X(x) = \varepsilon$ for $x \in (0, 1]$. Note that this distribution is single-peaked and has expected value $EX = \varepsilon/2$. Since TISR elicits the most likely events, $L^* = 0$. 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 elicited under TISR does not include the mean of X. ■

6.2 Inference on the Dispersion of Beliefs

The width of the interval for a given scoring rule depends on u and F_X. We show that the interval width of the interval increases when beliefs become more noisy in the following sense.

Definition 7 *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 $[a, b]$.

We consider noisiness to be an intuitive measure of uncertainty, since the uniform distribution can be interpreted as the case where the expert has no information. Note that under this notion of noisiness, unlike a mean preserving spread, the expected value typically changes when noise increases.

Proposition 8 *Assume $\gamma \geq 1/2$. If X' is noisier than X, then*

$$W^*(F_X, S_T, u) \leq W^*(F_{X'}, S_T, u).$$

Proposition 8 establishes that an elicitor can use the TISR to get insights into the degree of noisiness or dispersion of the beliefs of the expert. However, unless the preferences

of the expert are known, inference about dispersion will be confounded with inferences about the risk aversion of the expert.

Intuitively, one would expect that experts who are more risk averse will specify larger intervals, since they are more worried about getting a payoff of 0. This intuition can be formalized as follows. We say that $\tilde{u}$ is more risk averse than u if there is a concave function g such that $\tilde{u}(x) = g(u(x))$ for all x.

Proposition 9 *Assume $\gamma \geq 1/2$. If $\hat{u}$ is more risk averse than u then*

$$W^*(F_X, S_T, u) \leq W^*(F_X, S_T, \tilde{u}).$$

Proposition 9 tells us that a more risk averse expert will always specify a weakly larger width.[7]

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

The results from several experimental studies using the ISR (the untruncated version of the TISR) confirm these comparative statics.[8] In the experiment by Galbiati, Schlag, and Van der Weele (2011), average interval widths (measured within-subject) declined substantially in a treatment where uncertainty about the other player's actions was hypothesized to go down. Cettolin and Riedl (2011a,b) find a positive correlation between a measure of risk aversion and interval width, which is significant at 1% in Cettolin and Riedl (2011a).[9]

[7]The proof of Proposition 9 reveals that $[L^*(X, S_T, u), U^*(X, S_\gamma, u)] \subseteq [L^*(X, S_T, \hat{u}), U^*(X, S_T, \hat{u})]$.

[8]As remarked above, inferences and comparative statics from the ISR are the same as from TISR, but intervals may be larger than necessary.

[9]The results of Cettolin and Riedl are not reported in their papers, but confirmed by personal correspondence.

7 Extensions

In this section we discuss some of the assumptions that we have made above.

7.1 Multi-peaked Distributions

When a distribution may reasonably be expected to have more than one peak, it makes sense to give the expert the possibility to specify more than one interval. In the most general case one allows the expert to elicit any set, and W now equals the Lebsgue measure of this set. Whether this set is convex (the single interval case) or not does not affect our formal results, but it will affect the complexity of the elicitation procedure.

7.2 More Complicated Rules

Simple rules are advantageous in experimental practice, as they are easily presented to subjects. If one is willing to consider more complicated rules, one can extend TISR by adding a reward $\left(\frac{c}{1-W}\right)$ with an appropriately chosen constant $c > 0$, in case $x \notin [L, U]$. The resulting rule has coverage γ and dominates TISR with respect to the interval width for all F_X and all u.[10] It may be possible to devise even more precise rules by adding terms. We leave this issue to future research, but want to note the trade-offs involved between simplicity and precision.

8 Conclusion

Eliciting belief intervals is a much used practice, and a good way to gain a quick and intuitive understanding of both the events that the expert thinks likely to occur and the dispersion of an expert's beliefs. However, there is no established methodology or theory for incentivized elicitation of such intervals. Moreover, existing interval scoring rules lack some desirable properties. A particular drawback is the fact that one cannot infer which events the expert thinks are the most likely to occur. The 'Truncated Interval Scoring Rule' (TISR) proposed in this paper remedies these problems, and we believe its simplicity makes it an attractive belief elicitation method for experimentalists.

[10] Proofs are available as supplementary material at the authors' homepages.

The appeal of confidence intervals merits further work into interval scoring rules. On the empirical side, it will be necessary to compare the performance of these and other interval scoring rules. On the theoretical side, there are further questions about the trade-offs in designing interval scoring rules, some of which we have highlighted throughout the paper.

References

Aitchison, J., and I. Dunsmore (1968): "Linear-loss interval estimation of location and scale parameters," *Biometrika*, 55(1), 141–148.

Armantier, O., and N. Treich (2010): "Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging," *IDEI Working Paper*, 643.

Blavatskyy, P. R. (2008): "Betting on own knowledge: Experimental test of overconfidence," *Journal of Risk and Uncertainty*, 38(1), 39–49.

Bottazzi, G., G. Devetag, and F. Pancotto (2011): "Does Volatility Matter? Expectations of Price Return and Variability in and Asset Pricing Experiment," *Journal of Economic Behavior & Organization*, 77(2), 124–146.

Budescu, D. V., and N. Du (2007): "Coherence and Consistency of Investors' Probability Judgments," *Management Science*, 53(11), 1731–1744.

Casella, G., and J. Hwang (1991): "Evaluating confidence sets using loss functions," *Statistica Sinica*, 1, 159–173.

Cesarini, D., O. Sandewall, and M. Johannesson (2006): "Confidence interval estimation tasks and the economics of overconfidence," *Journal of Economic Behavior & Organization*, 61(3), 453–470.

Cettolin, E., and A. Riedl (2011a): "Fairness and Uncertainty," *Manuscript, Maastricht University.*

Cettolin, E., and A. M. Riedl (2011b): "Partial Coercion, Conditional Cooperation, and Self-Commitment in Voluntary Contributions to Public Goods," *Meteor Working Paper*, RM/11/041.

Galbiati, R., K. H. Schlag, and J. J. Van der Weele (2011): "Sanctions that Signal: an Experiment," *Working Paper, University of Vienna*, 1107.

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

Hamill, T., and D. Wilks (1995): "A Probabilistic Forecast Contest and the Difficulty in Assesssing Short-Range Forecast Uncertainty," *Weather and Forecasting*, 10.

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

Kirchler, E., and B. Maciejovsky (2002): "Simultaneous over- and underconfidence : Evidence from experimental asset markets," *Journal of Risk and Uncertainty*, pp. 1–24.

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

Moore, D. a., and P. J. Healy (2008): "The trouble with overconfidence.," *Psychological review*, 115(2), 502–17.

Offerman, T., J. Sonnemans, G. Van de Kuilen, and P. P. Wakker (2009): "A Truth Serum for Non-Bayesians," *Review of Economic Studies*, pp. 1–46.

Peeters, R., M. Vorsatz, and M. Walzl (2012): "Beliefs and truth-telling: A laboratory experiment," *University of Innsbruck Working Paper*, 17.

Riedl, A., and P. Smeets (2011): "Strategic and Non-Strategic Pro-Social Behavior in Financial Markets," *Manuscript, Maastricht University*.

Schlag, K. H., and J. J. van der Weele (2012): "Eliciting Probabilities, Means, Medians, Variances and Covariances without assuming Risk Neutrality," *Manuscript, Vienna University*.

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

Winkler, R. (1972): "A Decision-Theoretic Approach to Interval Estimation," *Journal of the American Statistical Association*, 67(337), 187–191.

Winkler, R., and A. Murphy (1979): "The use of probabilities in forecasts of maximum and minimum temperatures," *The Meteorological Magazine*, 108(1288), 317—-329.

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

Yaniv, I., and D. Foster (1997): "Precision and accuracy of judgmental estimation," *Journal of behavioral decision making*, 10, 21–32.

Appendix with Mathematical Proofs

We assume throughout that $u(0) = 0$, $a = 0$ and $b = 1$. Note that this can be done without loss of generality by appropriate rescaling of the scoring rule. If S is a scoring rule for $X \in \Delta[0,1]$ with coverage γ then $\bar{S}$ is a scoring rule for $X \in \Delta[a,b]$ with the same coverage if $\bar{S}(L,U,x) = S\left(\frac{L-a}{b-a}, \frac{U-a}{b-a}, \frac{x-a}{b-a}\right)$.

Proof of Proposition 1.
Rule of Schmalensee (1976). Given that the rule of Schmalensee (1976) depends on the midpoint as well as on the width, we define the variables B and R, so that R is the midpoint, $2B$ is the width, and the specified interval is $[R-B, R+B]$. Given upper hemi-continuity of the expected utility in the elicited parameters, we can limit attention in the proof to distributions that have no point masses.

The expected utility is given by

$$
\begin{aligned}
Eu\left(S\right) = & \int_{0}^{R-B} u\left(-\left(\frac{1-\gamma}{2}\right)2B-\left(R-x\right)-\left(R-B-x\right)\right)f\left(x\right)dx \\
& + \int_{R-B}^{R} u\left(-\left(\frac{1-\gamma}{2}\right)2B-\left(R-x\right)\right)f\left(x\right)dx \\
& + \int_{R}^{R+B} u\left(-\left(\frac{1-\gamma}{2}\right)2B-\left(x-R\right)\right)f\left(x\right)dx \\
& + \int_{R+B}^{1} u\left(-\left(\frac{1-\gamma}{2}\right)2B-\left(x-R\right)-\left(x-R-B\right)\right)f\left(x\right)dx.
\end{aligned}
$$

The first order condition is

$$
\begin{aligned}
\frac{d}{dB}Eu\left(S\right) = & \gamma\int_{0}^{R-B} u'\left(-(1-\gamma)B-\left(R-x\right)-\left(R-B-x\right)\right)f\left(x\right)dx \\
& -\left(1-\gamma\right)\int_{R-B}^{R} u'\left(-(1-\gamma)B-\left(R-x\right)\right)f\left(x\right)dx \\
& -\left(1-\gamma\right)\int_{R}^{R+B} u'\left(-(1-\gamma)B-\left(x-R\right)\right)f\left(x\right)dx \\
& +\gamma\int_{R+B}^{1} u'\left(-(1-\gamma)B-\left(x-R\right)-\left(x-R-B\right)\right)f\left(x\right)dx,
\end{aligned}
$$

where we suppressed the terms outside the integrals, which cancel out. Using the fact that $u'(\cdot)$ is decreasing, we obtain

$$
\begin{aligned}
\frac{d}{dB}Eu\left(S\right) \geq & \gamma\, u'\left(-(1-\gamma)B-B\right)\left(\int_{0}^{R-B} f\left(x\right)dx+\int_{R+B}^{1} f\left(x\right)dx\right) \\
& -\left(1-\gamma\right)u'\left(-(1-\gamma)B-B\right)\left(\int_{R-B}^{R} f\left(x\right)dx+\int_{R}^{R+B} f\left(x\right)dx\right) \\
= & \, u'\left(-(1-\gamma)B-B\right)\left(\gamma-\int_{R-B}^{R} f\left(x\right)dx-\int_{R}^{R+B} f\left(x\right)dx\right) \\
= & \, u'\left(-(1-\gamma)B-B\right)\left(\gamma-P\left(X\in\left[R-B,R+B\right]\right)\right).
\end{aligned}
$$

Thus, if $P\left(X\in\left[R-B,R+B\right]\right)<\gamma$, the first order condition is positive and the expert will want to expand the interval. This implies that the rule has coverage γ for all concave utility.

Rule of Winkler and Murphy (1979). The proof is analogous and omitted here for reasons of space. ■

Proof of Proposition 2. Consider distribution F_ε that has density $f(0) = \varepsilon$ and $f(x) = 1 - \varepsilon$ for $x \in (0, 1)$, where $\varepsilon < \frac{1-\gamma}{2}$.

Rule of Winkler and Murphy (1979). WM79 selects the $\frac{\gamma}{2}$ and $\frac{1+\gamma}{2}$ quantiles, so we have

$$\begin{aligned} \varepsilon + (1-\varepsilon) L^* &= \frac{1-\gamma}{2} \\ L^* &= \frac{1}{2}\left(\frac{1-2\varepsilon-\gamma}{1-\varepsilon}\right), \end{aligned}$$

and

$$\begin{aligned} \varepsilon + (1-\varepsilon) U^* &= \frac{1+\gamma}{2} \\ U^* &= \frac{1}{2}\left(\frac{1-2\varepsilon+\gamma}{1-\varepsilon}\right). \end{aligned}$$

As a consequence

$$\begin{aligned} U - L &= \frac{1}{2}\left(\frac{1-2\varepsilon+\gamma}{1-\varepsilon}\right) - \frac{1}{2}\left(\frac{1-2\varepsilon-\gamma}{1-\varepsilon}\right) \\ &= \frac{\gamma}{1-\varepsilon} \end{aligned}$$

This implies that $\sup_{F_\varepsilon}(U-L) \geq \frac{\gamma}{1-\varepsilon}$ for $\varepsilon < \frac{1-\gamma}{2}$, and $\sup_{F\in\Delta} \geq \frac{\gamma}{1-\frac{1-\gamma}{2}} = \gamma\left(\frac{2}{1+\gamma}\right)$. The fact that the rule does not attain minmax width follows from the fact that TISR attains a maximum width of $\gamma > \gamma\left(\frac{2}{1+\gamma}\right)$ (see Proposition 6).

Rule of Schmalensee (1976). Maximizing the expected utility with respect to the midpoint and the width (omitted here for reasons of space), yields the same solution for the lower and upper bound as WM79. ■

Proof of Proposition 3. 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 $u(S_T(L, U, X))$ is upper semi-continuous in L and U. Note that $u\left((1-(U-L))^{\frac{1-\gamma}{\gamma}}\right)$ 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 $\Pr(X \in [L,U]) = \Pr(X \leq U) - \Pr(X < L)$. Note that the cdf F_X of X 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 $\Pr(X \leq U) \leq \Pr(X \leq U_0) + \varepsilon/2$. Let $F_X^-(x) = P(X < x)$, which is left-continuous and non-increasing. This implies that $\Pr(X < L) = F_X^-(L)$ 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 $u(S_T(L,U,X))$ is upper semi-continuous. ■

Proof of Proposition 4. 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 separate the problem into the one of finding the best choice of L and U for given $W = w$ and the problem of how to find the best w. In step 3 we show that expected utility is increasing in w whenever $M(w) < \gamma$.

Step 1. Since F_X is monotonically increasing, it is differentiable almost everywhere (see e.g. Gordon 1994, p. 514).[11] Let f be its derivative when it exists and right continuous otherwise. So $f \geq 0$. Since X is single-peaked, there exists x_0 such that f is monotonically increasing for $x < x_0$ and monotonically decreasing for $x > x_0$ and any mass point of X must be equal to x_0. In particular, X has at most one mass point. Let $\xi = \Pr(X = x_0)$. Together, this implies that $F_X(x) = \int_0^x f(x)\,dx + \xi * 1_{\{x \geq x_0\}}$. 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. For each $w \in [0,\gamma]$ let $h(w) = (1-w)^{\frac{1-\gamma}{\gamma}}$ and let $M(w) = P(X \in [L^*(w), U^*(w)])$ where $(L^*(w), U^*(w)) \in \arg\max_{L,U:U-L=W} u(S_T(L,U,X))$. Thus M is increasing in w, hence differentiable almost everywhere.

Step 3. Consider $w \in [0,\gamma]$ such that M is differentiable at w. Then

$$\begin{aligned}\frac{d}{dw}\left(u(h(w))M(w)\right) &= M'(w)\,u(h(w)) + M(w)\,u'(h(w))\,h'(w) \\ &= M'(w)\,u(h(w)) - M(w)\,u'(h(w))\left(\frac{1-\gamma}{\gamma}\right)\left(\frac{h(w)}{1-w}\right). \quad (2)\end{aligned}$$

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

As u' is concave, $u'(z) \leq u(z)/z$ and hence

$$\frac{d}{dw}\left(u\left(h\left(w\right)\right)M\left(w\right)\right) \geq M'\left(w\right)u\left(h\left(w\right)\right) - M\left(w\right)u\left(h\left(w\right)\right)\left(\frac{1-\gamma}{\gamma}\right)\left(\frac{1}{1-w}\right).$$

Note that M is concave by single-peakedness of X. Hence, the incremental mass $M'(w)$ captured by increasing w is decreasing, so the mass $1 - M(w)$ not covered is at most equal to the marginal increase in mass $M'(w)$ due to enlargening w times the part of the parameter space not covered $1 - w$. In other words, $1 - M(w) \leq M'(w)(1-w)$. Thus

$$\begin{aligned}\frac{d}{dw}\left(u\left(h\left(w\right)\right)M\left(w\right)\right) &\geq \frac{1-M\left(w\right)}{1-w}u\left(h\left(w\right)\right) - M\left(w\right)u\left(h\left(w\right)\right)\left(\frac{1-\gamma}{\gamma}\right)\left(\frac{1}{1-w}\right)\\ &= \left(1-\frac{M\left(w\right)}{\gamma}\right)\frac{u\left(h\left(w\right)\right)}{1-w}.\end{aligned}$$

Hence we have shown that if w is such that $M'(w)$ exists and $M(w) < \gamma$ then

$$\frac{d}{dw}\left(u\left(h\left(w\right)\right)M\left(w\right)\right) > 0.$$

Therefore, $M^* \geq \gamma$. ■

Proof of Proposition 7. Assume that S_T is dominated within $\mathcal{S}_{simple}$. So there exists $S \in \mathcal{S}_{simple}$ with coverage γ that satisfies $W^*(F_X, S, u) \leq W^*(F_X, S_T, u)$ for all $F_X \in \Delta$ and all concave u with strict inequality for some F_X and some u. In the following we will show that S is identical to S_T up to a constant factor when $U - L \leq \gamma$. Since $W^*(F_X, S_T, u) \leq \gamma$ this then implies that $W^*(F_X, S, u) = W^*(F_X, S_T, u)$ for all X and u which contradicts the above and hence proves that S_T is undominated among the simple scoring rules.

Consider the class of random variables X_z indexed by z for $z \in [0, \gamma]$ where the underlying distribution F_{X_z} puts point mass $\frac{\gamma - z}{1-z}$ on $x = 0$ and has density $f_z(x) = \frac{1-\gamma}{1-z}$ for $x \in [0,1]$. So $F_{X_z}(z) = \gamma$. It follows from (2) that $L^*(F_{X_z}, S_T, Id) = 0$ and $U^*(F_{X_z}, S_T, Id) = z$ so $M^*(F_{X_z}, S_T, Id) = \gamma$ where $Id(x) \equiv x$. As S_T covers exactly γ of the mass of X_z, so must S, which means that $L^*(X_z, S, Id) = 0$ and $U^*(F_{X_z}, S, Id) = z$ and hence $W^*(F_{X_z}, S, Id) = z$ for all $z \in [0, \gamma]$. Let r_z be the function defined on $(0, \gamma]$ such that $r_z(U) = S(0, U, in)$ where $S(L, U, in) = S(L, U, x)$ for $x \in [L, U]$. Given $0 < z \leq \gamma$, the first order conditions imply $F_{X_z}(z) * r_z'(z) + f_z(z) r_z(z) = 0$ and

hence $\gamma r_z'(z) + \frac{1-\gamma}{1-z} r_z(z) = 0$. We solve this first order differential equation and obtain $r_z(z) = c * (1-z)^{(1-\gamma)/\gamma}$ for $z \in (0, \gamma]$ and some $c > 0$. It follows that $S(0, U, in) = c * S_T(0, U, in)$ for all $0 < U \leq \gamma$.

We now show that $S(L, U, in) = c * S_T(L, U, in)$ holds more generally for $U - L \leq \gamma$. Consider 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 $x \in [0, 1]$. Due to the constant density we can assume that $U^*(X_z, S, Id) = U^*(X_z, S_T, Id) = \gamma$. It follows, by replicating the above arguments, and defining $r_z(L) = S(L, \gamma, in)$, that $S(L, \gamma, in) = \bar{c} * S_T(L, \gamma, in)$ for some $\bar{c} > 0$. Combining this with our previous analysis, setting $L = 0$, shows that $c = \bar{c}$. Continuing this way one can show, tediously, that $S(L, U, in) = c * S_T(L, U, in)$ whenever $U - L \leq \gamma$ which completes the proof. ■

Proof of Proposition 8. Consider random variables X, Y and X_ε as in Definition 7. Let $[L_\varepsilon^*, U_\varepsilon^*]$ be the optimal interval selected under X_ε and let $W_\varepsilon^* = U_\varepsilon^* - L_\varepsilon^*$. Let $M_\varepsilon(w) = P(X_\varepsilon \in [L^*(w), U^*(w)])$ so $M_\varepsilon(w) = (1-\varepsilon) M_0(w) + \varepsilon w$. Assume that $\frac{d}{dw}(u(h) M_0) \geq 0$. As M_0 is concave in w, $M_0' \leq M_0/w$, it follows that $\frac{d}{dw}(u(h) M_0) = u'(h) h' M_0 + u(h) M_0' \leq \left(u'(h) h' + \frac{1}{w} u(h)\right) M_0$. Hence, $\frac{d}{dw}(u(h) M_\varepsilon) = (1-\varepsilon) \frac{d}{dw}(u(h) M_0) + \varepsilon \left(u'(h) h' + \frac{1}{w} u(h)\right) w \geq 0$. As $\gamma \geq 1/2$, S_T is single-peaked and hence $W_\varepsilon^* \geq W_0^*$. ■

Proof of Proposition 9. 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 $\hat{u}(x) = g(u(x))$. Using concavity of g we obtain

$$\frac{d}{dw}(\hat{u}(h) M) = g'(u(h)) u'(h) h' M + \hat{u} M' \geq \left(\frac{1}{u(h)} u'(h) h' M + M'\right) g(u(h)).$$

So if

$$\frac{d}{dw}(u(h) M) = u'(h) h' M + u(h) M' \geq 0$$

then $\frac{d}{dw}(\hat{u}(h) M) \geq 0$ which completes the proof. ■