Proper Scoring Rules and Risk Aversion

Alexander Peysakhovich and Mikkel Plagborg-Møller*

March 9, 2012

Abstract

The literature on proper scoring rules has mostly studied the case of risk neutral agents. We analytically investigate how risk averse, expected utility maximizing forecasters behave when presented with risk neutral proper scoring rules. If the state variable is binary, risk averse agents shade their reports toward saying that the states are equally likely. In the non-binary case reported probabilities are compressed relative to truthtelling. We show the implications of our results for the use of elicited probabilities as inputs to decision-making and find that naive elicitors may violate first-order stochastic dominance. Possible resolutions of these problems are presented, including an estimator for the mean population belief when the distribution of risk attitudes is known. Finally, we discuss the relevance of our results to recent work in experimental economics.

1 Introduction

The problem of incentivizing agents to truthfully reveal their subjective probability estimates of given events is interesting from both a theoretical and an applied viewpoint.[1] One possible solution to this problem, proposed by Brier (1950), Good (1952), McCarthy (1956) and Savage (1971), among others, is the use of proper scoring rules. Scoring rules are mechanisms

*Dept. of Economics, Harvard University. {apeysakh, plagborg}@fas.harvard.edu. We thank Yiling Chen, Drew Fudenberg, Scott Kominers, Al Roth and seminar participants at Harvard University for insightful comments (in light of our results, we chose not to use proper scoring rules to elicit comments).

[1] For the less cynically inclined reader we quote Prelec (2004): "[. . .] I do not suggest that people are deceitful or unwilling to provide information without explicit financial payoffs. The concern, rather, is that the absence of external criteria can promote self-deception and false confidence even among the well-intentioned."

that can be applied in situations where uncertainty about a future event is eventually resolved. Suppose that we are interested in eliciting the subjective probability that an economic forecaster assigns to the federal funds rate exceeding 1% next Monday. The forecaster reports a probability q of the event occurring and on Monday we pay her $s_0(q)$ if the interest rate is at or above 1% and $s_1(q)$ if it is not. A proper scoring rule is such a function whose expectation with respect to the forecaster's true subjective probability p is maximized by reporting $q = p$. Proper scoring rules and their properties are well studied in the literature (Gneiting and Raftery, 2007, provide a very general overview), and they have been widely used in several applied fields ranging from accounting to medicine (see the introduction in Offerman et al., 2009, for a comprehensive list).

Though proper scoring rules induce expected *payoff* maximizing (i.e., risk neutral) agents to report truthfully, the case of risk averse forecasters has not been given a full analysis in the literature. Savage (1971, Section 3) provides an informal treatment of the topic, while Bickel (2007) investigates the numerical performance of certain commonly used scoring rules in the face of risk aversion. Kadane and Winkler (1988, Section 3) study the quadratic scoring rule when forecasters are risk averse, while Offerman et al. (2009) expand the analysis of this particular scoring rule to non-expected utility theory. In this paper we adopt a systematic, analytic approach to characterizing the behavior of risk averse, expected utility maximizing forecasters who are faced with a proper scoring rule. Our main conclusion is that truthful reporting is no longer incentive compatible in the generic case. In fact, forecasters distort their subjective probabilities in a particular way: When the state variable is binary, they report a distribution closer to 50-50 than their true opinion, and shading increases as agents become more risk averse. Extending our analysis to the case of finite state spaces with more than two outcomes, we show that optimal reports are compressions of true subjective beliefs but not necessarily uniformly shaded toward the uniform distribution.

Proper scoring rules offer an appealing framework for extracting subjective probability estimates from an expert for decision-making purposes. However, we show that a center that naively makes decisions based on reported probabilities from risk averse agents may violate first-order stochastic dominance with respect to the true subjective beliefs. Manski (2004) suggests that empirical economists begin to measure expectations directly and some experimental economists have turned to proper scoring rules to incentivize truthful revelation of beliefs (McKelvey and Page, 1990; Offerman et al., 1996; Costa-Gomes and Weizsäcker, 2008). Since subjects exhibit sig-

nificant risk aversion even when faced with small gambles (Holt and Laury, 2002, 2005), we argue that our results call for caution when evaluating experimental data that rely on scoring rules as a tool for belief elicitation. In this vein, we offer an alternative interpretation of the results in Nyarko and Schotter (2002) and Palfrey and Wang (2009). Additionally, we briefly survey existing methods of addressing the reporting problem under risk aversion, including payment via a binary lottery procedure and adjustment of the scoring rule, as well as their limitations. Instead of eliciting individual beliefs, a center may prefer to extract aggregate predictions from a group of risk-averse experts. We propose a mechanism which acts as a consistent estimator of the mean population belief when the population distribution of risk attitudes is known. Numerical simulations suggest that the estimator has good finite sample and robustness properties.

The paper proceeds as follows. Section 2 analyzes optimal reporting in the binary and multi-outcome cases. Section 3 applies the theoretical predictions to the literature on expectation elicitation. Finally, Section 4 concludes and suggests possible areas of future research. Proofs are relegated to the appendix.

2 Optimal reporting under risk aversion

The question raised in this section is how a risk averse forecaster should optimally respond to a scoring rule that is designed to induce truthtelling under the assumption of risk neutrality. After laying out a few basic definitions and assumptions, we treat the special but illustrative case of binary forecast variables. Subsequently, the more general setting with finite outcome spaces is discussed.

2.1 Basic model

Consider a stochastic *state* variable X that takes values in a finite outcome space $\Omega = \{1, 2, \ldots, n\}$ with $n \geq 2$ possible outcomes. A probability distribution over the outcomes is a vector contained in the n-simplex $\Delta_n = \{\mathbf{p} \in \mathbb{R}_+ \mid \sum_{i=1}^n p_i = 1\}$. Sometimes we will restrict attention to the open n-simplex $\Delta_n^\circ = \{\mathbf{p} \in \mathbb{R}_+ \mid \sum_{i=1}^n p_i = 1, p_i > 0 \text{ for all } i\}$.

A *center* is interested in collecting accurate probabilistic predictions of X. Consider a representative expected utility maximizing *forecaster* (also referred to as the *expert*) with subjective probability estimate $\mathbf{p} \in \Delta_n$. Her report is denoted $\mathbf{q} \in \Delta_n$, with *truthful reporting* corresponding to $\mathbf{q} = \mathbf{p}$.

Definition 1. *A* scoring rule *is a vector-valued function* $\mathbf{s}\colon D \to \mathbb{R}^n$*, where* D *is either* Δ_n *or* Δ_n°*. The scoring rule is* strictly proper *if for all* $\mathbf{p} \in D$ *it holds that*

$$\mathbf{p} = \operatorname*{argmax}_{\mathbf{q}\in D} \sum_{i=1}^{n} p_i s_i(\mathbf{q}). \tag{1}$$

Hence, a scoring rule is a function linking monetary rewards $s_i(\mathbf{q})$ in each of the states $i = 1, \ldots, n$ to the issued probability report $\mathbf{q}$. The scoring rule is said to be strictly proper if it incentivizes an expected payoff maximizing (i.e., risk neutral expected utility maximizing) forecaster to reveal her true subjective probability $\mathbf{p}$ (see Gneiting and Raftery, 2007, for a general treatment). Well-known examples of strictly proper scoring rules include the logarithmic $s_i(\mathbf{q}) = \log q_i$, the quadratic $s_i(\mathbf{q}) = 2q_i - ||\mathbf{q}||^2$ and the spherical $s_i(\mathbf{q}) = q_i/||\mathbf{q}||$ scoring rules ($||\cdot||$ denotes the Euclidean norm).

Apart from strict propriety, we shall impose a number of weak assumptions on the scoring rule. They are necessary when applying standard optimization theorems to the forecaster's utility maximization problem.

Assumption 1 (Neutrality)**.** *For all* $\mathbf{q} \in D$ *and all* n*-permutations* $\sigma(\cdot)$ *it holds that* $\mathbf{s}(\sigma(\mathbf{q})) = \sigma(\mathbf{s}(\mathbf{q}))$*.*

Neutrality amounts to requiring that the scoring rule is invariant under relabeling of the outcomes.[2]

Assumption 2 (Boundedness above)**.** *Component function* $s_i(\cdot)$ *is bounded above for all* $i = 1, \ldots, n$*.*

Boundedness below is not needed for the analysis, so the logarithmic scoring rule does not present any challenges to the generality.

Assumption 3 (Extended continuity)**.** *The scoring rule* $\mathbf{s}(\cdot)$ *is continuous on* D*, and for all* $\mathbf{q} \in \Delta_n$ *(not just* D*) and any sequence* $\{\mathbf{q}^k\}$ *with* $\lim_{k\to\infty} \mathbf{q}^k = \mathbf{q}$ *it holds that* $\lim_{k\to\infty} \mathbf{s}(\mathbf{q}^k)$ *exists and is independent of the particular choice of* $\{\mathbf{q}^k\}$*.*

Extended continuity is equivalent to continuity if $D = \Delta_n$ and only represents a slight strengthening in the case $D = \Delta_n^\circ$. The assumption is

[2]In particular, the range $R = \{s_i(\tilde{\mathbf{p}}) \mid \tilde{\mathbf{p}} \in D\}$ must be independent of the index i. Also, if $i, j, k = 1, \ldots, n$ are distinct indices, $s_i(\sigma_{jk}(\mathbf{q})) = s_i(\mathbf{q})$, where $\sigma_{jk}(\cdot)$ denotes the permutation that switches the j-th and k-th coordinates. Hence, the payout in state i is determined by q_i as well as the magnitudes, but not the ordering, of the remaining components of $\mathbf{q}$.

needed to rule out pathological boundary cases when proving existence of an optimal report. Clearly, each of the three above-mentioned specific scoring rules is neutral, bounded above and extended continuous. In addition, any affine transformation $\tilde{s}_i(\mathbf{q}) = a + bs_i(\mathbf{q})$ (where a and $b > 0$ do not depend on i) preserves these properties. The oft-used Brier score (Brier, 1950) is an affine transformation of the quadratic scoring rule.

We consider a forecaster whose utility function $u\colon \mathrm{cl}(R) \to \mathbb{R}$ of wealth is strictly increasing, twice differentiable and strictly concave, meaning that the forecaster is risk averse. Here $\mathrm{cl}(R)$ denotes the closure of the range of $\mathbf{s}(\cdot)$. The form of the utility function is unknown to the center. Let

$$S_{\mathbf{p}}(\mathbf{q}) = \sum_{i=1}^{n} p_i u(s_i(\mathbf{q}))$$

denote the expected utility of reporting $\mathbf{q}$ when the subjective belief is $\mathbf{p}$. The rest of Section 2 will be concerned with characterizing utility-maximizing reports as a function of the subjective probabilities.

Definition 2. *Let the domain of the scoring rule $\mathbf{s}(\cdot)$ be D. The* optimal report correspondence $\mathbf{q}\colon D \rightrightarrows D$ *is defined by*

$$\mathbf{q}(\mathbf{p}) = \left\{ \tilde{\mathbf{q}} \in D \;\middle|\; S_{\mathbf{p}}(\tilde{\mathbf{q}}) = \sup_{\hat{\mathbf{q}} \in D} S_{\mathbf{p}}(\hat{\mathbf{q}}) \right\}$$

for all $\mathbf{p} \in D$. If $\mathbf{q}(\mathbf{p})$ is a singleton for each $\mathbf{p} \in D$, we say that the optimal report function *exists and we treat it as a function $\mathbf{q}\colon D \to D$.*

2.2 Binary outcome space

To provide some intuition and highlight a few special results, we focus on a binary state space $\Omega = \{0, 1\}$. In this case, neutrality of the scoring rule is equivalent to the condition that $s_0(q, 1-q) = s_1(1-q, q)$ for all $q \in D$, where the domain D is either $[0, 1]$ or $(0, 1)$. Hence, when working with neutral scoring rules on a binary space, it is convenient to drop the subscripts and write $s(q)$ (resp., $s(1-q)$) for the state-0 (resp., state-1) payoff. Similarly, we write $q(p)$ for the optimal report of the state-0 probability.

Lemma 1. *Assume $s(\cdot)$ is a strictly proper and continuously differentiable scoring rule defined on D, where the latter is either $[0, 1]$ or $(0, 1)$. If $s(\cdot)$ satisfies assumptions 1–3, the optimal report function $q(\cdot)$ exists. It is continuous, increasing and symmetric, i.e., $q(p) = 1 - q(1-p)$ for all $p \in D$.*

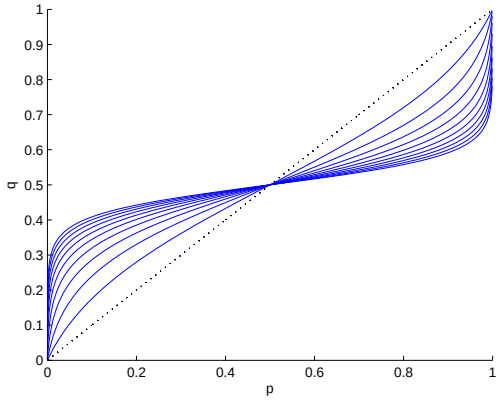

Figure 1: Plots of $q(p)$ for CARA utility functions $u_\alpha(x) = -\exp(-\alpha x)$ with absolute risk-aversion parameters $\alpha = 0.5, 1.0, 1.5, \ldots, 5.0$ and a quadratic scoring rule $s(q) = 2q - q^2 - (1-q)^2$. The dotted line is the 45 degree line. Numerical optimization was carried out in Matlab using `fminbnd`.

We shall now state the main results in the binary case. The reader may refer to Figure 1 for an illustration.

Proposition 1. *Assume that the scoring rule satisfies the assumptions in Lemma 1. The optimal report function has the following properties.*

(i) $1/2 \leq q(p) \leq p$ *for all* $p \in [1/2, 1] \cap D$. *The first equality only holds at* $p = 1/2$, *and the second only at* $p = 1/2$ *and* $p = 1$ *(if applicable).*

(ii) If $u(\cdot)$ *and* $\tilde{u}(\cdot)$ *are two utility functions satisfying the assumptions in the preamble, and the coefficient of absolute risk aversion of* $\tilde{u}(\cdot)$ *is uniformly strictly greater than that of* $u(\cdot)$, *then* $q_{\tilde{u}}(p) < q_u(p)$ *for all* $p \in (1/2, 1)$.

Note that due to symmetry, any properties of $q(\cdot)$ on $(1/2, 1)$ translate readily into properties on $(0, 1/2)$.[3]

As a closed-form example, a little calculus will show that for CARA utility $u(x) = -\exp(-\alpha x)$ and the logarithmic scoring rule $s(q) = a + b \log q$,

[3] Although Figure 1 suggests that the optimal report function is concave on $(0, 1/2)$, this regularity does not follow from our conditions on the utility function and scoring rule. One may construct a counterexample involving a utility function that has large curvature for small scoring rule values and small curvature for large values, thus inducing substantial shading for small p and close-to-truthtelling for larger p.

the optimal report function is given by $q(0) = 0$ and

$$q(p) = \frac{1}{1 + \left(\frac{1-p}{p}\right)^{1/(1+\alpha b)}} \quad \text{for } p \in (0, 1].$$

We see that in this special case the adverse effects of risk aversion can be mitigated by lowering b. More generally, property *(ii)* of Proposition 1 implies the following corollary.

Corollary 1. *Suppose the scoring rule $s(\cdot)$ satisfies the assumptions in Lemma 1.*[4] *Let the forecaster's belief $p \in (0, 1)$ satisfy $p \neq 1/2$.*

(i) Consider scaling the scoring rule to $\tilde{s}(q) = bs(q)$, where $b > 0$. If $u(\cdot)$ exhibits strictly increasing (resp., decreasing) relative *risk aversion, the amount $|q(p)-p|$ of shading is strictly increasing (resp., decreasing) in b. If $u(\cdot)$ belongs to the constant relative risk aversion family, the optimal report is invariant under scalings of the scoring rule.*

(ii) Consider shifting the scoring rule to $\tilde{s}(q) = s(q) + a$, where $a \in \mathbb{R}$. If $u(\cdot)$ exhibits strictly increasing (resp., decreasing) absolute *risk aversion, the amount $|q(p) - p|$ of shading is strictly increasing (resp., decreasing) in a. If $u(\cdot)$ belongs to the constant absolute risk aversion family, the optimal report is invariant under shifts of the scoring rule.*

If the forecaster is risk neutral, any affine transformation of a strictly proper scoring rule is still strictly proper. As the corollary demonstrates, under risk aversion the optimal report function is in general sensitive to a shift or scaling of the scoring rule. In addition to implying testable predictions of our model, Corollary 1 highlights the inherent tradeoffs when transforming the scoring rule.[5]

2.3 Finite outcome space

With a general finite outcome space $\Omega = \{1, 2, \ldots, n\}$ we need to introduce a little extra terminology before stating the main result. While the additional restrictions are not quite as intuitive as the preceding assumptions, they are easy to check and weak enough that we avoid losing much generality.

[4] The domain of the utility function must be extended to accommodate the transformations of the scoring rule mentioned below.

[5] For example, Pennock (2006) interprets the multiplicative constant b as a liquidity parameter for an automated marker maker who subsidizes trades according to the logarithmic scoring rule. Our results indicate that increased market liquidity may bring about more severe shading, depending on the specific risk attitudes of the investors.

Assumption 4 (Semi-strict quasiconcavity)**.** *For all $i = 1, \ldots, n$, $t \in \mathbb{R}$, $\alpha \in (0,1)$ and $\mathbf{q}', \mathbf{q}'' \in \Delta_n^\circ$ it holds that $s_i(\mathbf{q}') \geq t$ and $s_i(\mathbf{q}'') \geq t$ imply $s_i(\alpha\mathbf{q}' + (1-\alpha)\mathbf{q}'') \geq t$, with strict inequality whenever $q_i' \neq q_i''$.*

Semi-strict quasiconcavity requires that each of the n component functions $s_i(\mathbf{q})$ of a scoring rule is quasiconcave. Furthermore, when taking the convex combination of two different probability reports, some of the component functions must exhibit strict quasiconcavity (for that parameter choice). The semi-strict quasiconcavity assumption ensures that a risk averse forecaster has a unique optimal report.

The last assumption we shall need is purely technical.

Assumption 5 (Lagrange sufficiency)**.** *The scoring rule $\mathbf{s}(\cdot)$ is continuously differentiable and the Lagrangian $\mathcal{L}_\mathbf{p} \colon \Delta_n^\circ \times \mathbb{R} \to \mathbb{R}$ given by*

$$\mathcal{L}_\mathbf{p}(\mathbf{q}, \mu) = \sum_{i=1}^n p_i s_i(\mathbf{q}) - \mu \left(\sum_{i=1}^n q_i - 1 \right),$$

for $\mathbf{p}, \mathbf{q} \in \Delta_n^\circ$ and $\mu \in \mathbb{R}$, has a unique stationary point of the form $(\mathbf{p}, \mu_\mathbf{p}^*)$ for every $\mathbf{p}$, where $\mu_\mathbf{p}^*$ is a constant that may depend on $\mathbf{p}$.

The condition simply requires that the Lagrange first-order conditions associated with the risk neutral optimization problem (1) are sufficient for optimality. Since the objective function is strictly quasiconcave on Δ_n° if the scoring rule is semi-strictly quasiconcave, the latter property implies Lagrange sufficiency whenever the multiplier $\mu_\mathbf{p}^*$ at the optimum is nonzero (Mas-Colell et al., 1995, pp. 961–962). If the multiplier vanishes at the optimum, a stronger restriction on the curvature of the component functions $s_i(\cdot)$, such as concavity, is needed to make the first-order conditions sufficient.

It is easy to see that the logarithmic, quadratic and spherical scoring rules are each semi-strictly quasiconcave[6] and Lagrange sufficient.[7]

[6]For the quadratic scoring rule, each $s_i(\cdot)$ is strictly concave. The component functions in the logarithmic rule are each concave, and $s_i(\cdot)$ exhibits strict concavity whenever $q_i' \neq q_i''$. The component functions in the spherical scoring rule are each strictly quasiconcave: If $s_i(\mathbf{q}') \geq t$ and $s_i(\mathbf{q}'') \geq t$, we have $||\mathbf{q}'|| \leq q_i'/t$ and $||\mathbf{q}''|| \leq q_i''/t$, so

$$s_i(\alpha\mathbf{q}' + (1-\alpha)\mathbf{q}'') > \frac{\alpha q_i' + (1-\alpha) q_i''}{\alpha||\mathbf{q}'|| + (1-\alpha)||\mathbf{q}''||} \geq \frac{\alpha q_i' + (1-\alpha) q_i''}{\alpha q_i'/t + (1-\alpha) q_i''/t} = t.$$

[7]For any $\mathbf{p} \in \Delta_n^\circ$ the Lagrange multiplier at the optimum in (1) is nonzero when the scoring rule is logarithmic or spherical. It is zero for the quadratic rule (which may be interpreted as a consequence of Selten's (1998) extension axiom), but since the component functions are concave in this case, Lagrange sufficiency still holds.

Lemma 2. *Let $\mathbf{s}(\cdot)$ be a strictly proper scoring rule defined on the domain D, which is either Δ_n or Δ_n°. If $\mathbf{s}(\cdot)$ satisfies assumptions 1–5, the optimal report function exists. It is continuous and symmetric, i.e., $\mathbf{q}(\sigma(\mathbf{p})) = \sigma(\mathbf{q}(\mathbf{p}))$ for all n-permutations $\sigma(\cdot)$. In the case $D = \Delta_n$, $q_i(\mathbf{p}) = 0$ if and only if $p_i = 0$, and $q_i(\mathbf{p}) = 1$ if and only if $p_i = 1$.*

Symmetry of $\mathbf{q}(\cdot)$ readily implies that truthtelling is optimal at the uniform belief: $\mathbf{q}(\mathbf{e}/n) = \mathbf{e}/n$, where $\mathbf{e} = (1, 1, \ldots, 1)$. However, the following result, which is a partial generalization of Proposition 1, shows that truthtelling is suboptimal at all other non-degenerate beliefs.

Proposition 2. *Assume that the scoring rule satisfies the assumptions in Lemma 2. For all $i, j = 1, \ldots, n$ such that $p_i > p_j$ it holds that*

$$\frac{p_i}{p_j} > \frac{q_i(\mathbf{p})}{q_j(\mathbf{p})} > 1.$$

It is interesting to note that, contrary to what one might first surmise, risk aversion does not necessitate that the forecaster shade each coordinate of her report toward the uniform distribution $q_i = 1/n$.[8] Instead, risk aversion brings about a compression of the reported probability vector relative to the subjective belief vector. The economic intuition is that decreasing marginal utility causes forecasters to hedge their bets by moving payoff from high payoff, high probability states to low payoff, low probability states. The proof of the proposition establishes that, for all $i = 1, \ldots, n$,

$$q_i(\mathbf{p}) = \frac{p_i u'(s_i(\mathbf{q}(\mathbf{p})))}{\sum_{j=1}^n p_j u'(s_j(\mathbf{q}(\mathbf{p})))}.$$

The conclusion that risk averse probabilities are given by a weighted average of "true" probabilities, with weights determined by the marginal utility of wealth in the different states, is a familiar one in economics (Jackwerth, 2000).

3 Applications

Manski (2004) surveys the growing literature on elicitation of subjective probabilities in economics. In arguing the importance of learning agents' private beliefs, he mentions scoring rules as a potent way of incentivizing accurate reporting. In this section we apply our theoretical results to the problems of eliciting individual beliefs and surveying groups of agents.

[8] For example, if $\mathbf{p} = (0.3, 0.05, 0.65)$, $s_i(\mathbf{q}) = \log q_i$ and $u(x) = -\exp(-x)$, the optimal report has $q_1(\mathbf{p}) \approx 0.3472 > 1/3$.

3.1 Elicitation of individual beliefs

There are at least two reasons why it is of interest to elicit subjective probabilities from individuals. First, they can serve as inputs in decision-making mechanisms when a limited number of experts are available. Second, transforming private beliefs into observables in a lab setting allows for more powerful tests of economic theories.

An important normative criterion for decision-making under uncertainty is *first-order stochastic dominance*.[9] In light of this criterion the type of misreporting discussed in Section 2 creates hazards to a naive center who simply uses reported probabilities as inputs to decision-making. Fix a set of states of the world Ω with more than 2 possible states. Let $\mathcal{F}$ denote the set of *simple acts* (Savage, 1972) over Ω, so that each $f \in \mathcal{F}$ is a map $f\colon \Omega \to \mathbb{R}$. In other words, the choice of an act determines the payoff to the decision-maker in each state of the world. The combination of an act $f \in \mathcal{F}$ and a probability distribution $\mathbf{p} \in \Delta_n$ induces a lottery $(\mathbf{p}, f) \in \mathcal{M}(\mathbb{R})$, where $\mathcal{M}(\mathbb{R})$ is the set of probability distributions over $\mathbb{R}$.

Proposition 3. *Let the scoring rule and utility function of the forecaster satisfy the assumptions of Proposition 2. Suppose that a naive center, who maximizes the expectation of a strictly increasing utility function of wealth, uses reported probabilities* $\mathbf{q}(\mathbf{p})$ *to evaluate choices between lotteries. Then there exists probability distributions* $\mathbf{p}, \tilde{\mathbf{p}} \in \Delta_n$ *and an act* $f \in \mathcal{F}$ *such that the lottery* $(\mathbf{p}, f)$ *first-order stochastically dominates* $(\tilde{\mathbf{p}}, f)$ *but the center ranks* $(\mathbf{q}(\tilde{\mathbf{p}}), f)$ *over* $(\mathbf{q}(\mathbf{p}), f)$*.*

Intuition may lead one to think that the use of a proper scoring rule on risk averse agents would simply cause the center to make decisions in a more risk averse manner. However, the proposition shows that in fact naivety can result in seriously suboptimal decision-making. Consider a firm employing a proper scoring rule to elicit the beliefs of a forecaster regarding a project which may succeed to varying degrees. The result above states that the firm may choose to undertake the project when the forecaster has beliefs $\tilde{\mathbf{p}}$ and not when she has beliefs $\mathbf{p}$, even though the latter beliefs represent a more optimistic assessment of the prospects of the project.[10]

Scoring rules are interesting for purposes other than decision-making. Expectations form an important component of economic theories of behav-

[9] For finite probability spaces, a lottery L with probabilities $\mathbf{p}$ and associated payoffs $\mathbf{x}$ is said to first-order stochastically dominate another lottery $\tilde{L}$ with probabilities $\tilde{\mathbf{p}}$ and payoffs $\tilde{\mathbf{x}}$ if for all z it holds that $\sum_{i\colon x_i<z} p_i \leq \sum_{i\colon \tilde{x}_i<z} \tilde{p}_i$.

[10] Note that given the monotonicity result of Proposition 1, the first-order stochastic dominance result fails when we restrict to binary state spaces.

ior. In recent years both theorists and experimental economists have turned to building and testing theories of expectation formation. Here researchers have begun to apply proper scoring rules to turn beliefs into observables.[11] We argue that our theoretical results may cast new light on existing experimental work.

In order to test a theory of learning in games, Nyarko and Schotter (2002, henceforth NS) elicit the beliefs of subjects as they repeatedly play 2×2 matrix games. Using the quadratic scoring rule, they find that subjects' beliefs are generally very extreme, volatile and can be used to predict play via best responses approximately 75% of the time. Palfrey and Wang (2009, henceforth PW) take the data from the NS experiment and incentivize new subjects to predict the next play in a NS sequence using two proper scoring rules, the logarithmic[12] and quadratic, and one improper rule, the linear $s(q) = a + bq$. They find that the quadratic and logarithmic scoring rules induce different distributions of reports from subjects, and reports are most dispersed under the linear scoring rule. Additionally, they argue that proper scoring rules induce forecasts which seem to be more calibrated to empirical frequencies than the linear scoring rule. Our Proposition 1 suggests that another explanation for the results in NS and PW is that players are actually very poorly calibrated and overreact highly to new information due to recency effects, but the natural shading brought about by risk aversion conceals some of the overreaction. As an example of this logic, observe that if subjects behave according to the logarithmic utility function, the linear scoring rule actually induces truthful reporting. While our concerns are motivated purely by theory, they do indicate that researchers need to be careful when drawing inference based on scoring rules.

How might practitioners mitigate the problems caused by risk aversion? One method is to adjust the scoring rule itself to regain its incentive compatibility properties (Savage, 1971, refers to this as "paying in utiles"). If the preferences of the forecaster are known and described by a well-behaved utility function $u(\cdot)$, it is easy to see that a risk neutral proper scoring rule $s(\cdot)$ will be proper in the risk averse case when transformed as $u^{-1}(s(\cdot))$. However, this adjustment depends on a precise knowledge of the forecaster's utility function, whereas most experimental protocols are designed to calibrate within a specific family (e.g. CARA utility). Additionally, individual risk aversion is always measured with noise and can even be affected by

[11] There is some debate in the literature about whether this elicitation actually influences actions. For example, Costa-Gomes and Weizsäcker (2008) argue that elicitation changes behavior toward play that is better described by belief learning.

[12] To counteract unbounded losses, PW restrict reports to the range $[0.1, 0.9]$.

current emotional state (Lerner and Keltner, 2001). Another way to alter the mechanism to encourage truthful elicitation is the use of a binary lottery procedure where a prize is fixed and the payoffs of the scoring rule are lottery tickets for that prize (Roth and Malouf, 1979). Since expected utility is linear in probabilities, this should, in theory, produce risk neutral decisions.[13] Evidence on the effectiveness of binary lotteries is mixed (Kagel and Roth, 1997; Selten et al., 1999). Furthermore, it is impossible to implement the procedure with scoring rules, such as the logarithmic, that are unbounded below without restricting the domain of probabilities that forecasters are allowed to report.

Another practical question is the proper choice of event space for the forecasters.[14] The demand for a product is a nearly continuous variable but a firm may be interested only in the probability that the demand will exceed a certain threshold. When risk-neutral forecasters are presented with coarsenings of an underlying event space, their forecasts for a compound event will necessarily be additive (by strict propriety). However, this is not generally the case for risk-averse forecasters. Consider, the simple example of an event space $\Omega = \{1, 2, 3\}$. A risk-averse forecaster with uniform subjective probability will report truthfully if asked about the finest possible partition, but if it is coarsened to $\{1, \text{not } 1\}$, he will distort his report. Conversely, a forecaster with subjective probability $\mathbf{p} = (1/2, 1/4, 1/4)$ will distort if asked about the full event space but will report truthfully in the above-mentioned coarsening. Thus without further assumptions there is no systematic answer as to whether coarser or finer partitions of the full event space are better.[15]

3.2 Elicitation of group beliefs

In many cases centers may be interested in eliciting aggregate information from a group of agents rather than a single forecaster. The literature has considered three ways in which scoring rules may be used to harness "the wisdom of the crowd" when forecasters are risk neutral. First, the center can elicit everyone's beliefs and compute the average. Second, the payoff

[13] The original intuition for this procedure comes from Smith (1961) who describes how probability can be seen as a currency using the metaphor of hiding a diamond in a big pile of beeswax. In the early days of experimental economics this proved to be a quite difficult design to implement in the lab and so lottery tickets were used instead (Al Roth, private communication).

[14] Savage (1971) briefly mentions this issue in the context of multiple choice exams.

[15] While we do not pursue the idea here, it is possible that additional restrictions on the utility functions of the experts could yield experimentally testable predictions about subadditivity or superadditivity of reported probabilities.

structure of scoring rules may be incorporated into a prediction market setting (Hanson, 2007). Third, scoring rules can be combined with techniques from games of incomplete information to exploit the power of peer prediction (Miller et al., 2005; Prelec, 2004). Because these methods rely on the incentive compatibility of proper scoring rules, Propositions 2 and 3 indicate that caution must be applied in adjusting and interpreting group forecasts. As it may be difficult to elicit the risk attitude of each individual in a large group of agents, we propose a mechanism for this situation which only requires knowledge of the distribution of risk aversion in the population.

In the following we shall focus on the case with a binary state variable, but the results may readily be generalized. Suppose that forecasters have utility functions from a family that can each be described by a single parameter γ. Fixing some γ for forecaster i, the results from Section 2.2 yield that if the forecaster perfectly optimizes given the scoring rule and her true subjective probability, there exists a well-behaved function $q_\gamma \colon [0,1] \to [0,1]$ mapping subjective probabilities to shaded reports. Define $q_\gamma^{-1}(\cdot)$ to be the inverse map, i.e., $q_\gamma^{-1}(q_\gamma(p)) = p$ for all $p \in [0,1]$.

Proposition 4. *Let the number of forecasters be N. Suppose they each have a risk aversion parameter $\gamma_i \sim G(\gamma_i)$ and subjective belief $p_i \sim F(p_i)$, where $F(\cdot)$ and $G(\cdot)$ are arbitrary distributions on $[0,1]$ and $\Gamma \subset \mathbb{R}$, respectively. All pairs of risk aversion parameters and subjective beliefs are i.i.d., and these variables are not known to the center. Suppose the center knows $G(\cdot)$. Each forecaster releases a report $\tilde{q}_i$ that is optimal given γ_i and p_i, i.e., $\tilde{q}_i = q_{\gamma_i}(p_i)$. Then the estimator*

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} \int_\Gamma q_\gamma^{-1}(\tilde{q}_i) dG(\gamma)$$

is consistent for the mean population belief: $\hat{p} \xrightarrow{a.s.} E[p_1]$ as $N \to \infty$.

As is mentioned in the proof, Proposition 4 may be adapted to estimation of other moments of the belief distribution $F(\cdot)$, although implementation of the estimator requires a considerable amount of numerical work.[16]

The proposition assumes that forecasters report optimally according to the $q_\gamma(\cdot)$ function. Observe that if the population of forecasters as a whole were in fact more honest than implied by utility maximization, they would

[16] First of all, it will typically not be possible to solve for a closed-form expression for $q_\gamma^{-1}(\cdot)$. Second, the integral must be evaluated numerically, and for each of the iterations a new call to $q_\gamma^{-1}(\cdot)$ must be made.

announce less shaded reports, which would thus resemble those from utility maximizers with a lower degree of risk aversion. Hence, one may accommodate "excess honesty" by appropriately shifting the $G(\cdot)$ distribution. As detailed in Section A.8 of the appendix, the estimator appears to have good finite sample and robustness properties.

4 Conclusion

We demonstrated that when faced with a risk neutral proper scoring rule, truthtelling is generically not incentive compatible for risk averse forecasters. In the binary case forecasters shade their reports toward the equally-likely prediction. In the multiple outcome case forecasters compress their reports relative to truthtelling. These results are of interest to those who wish to apply proper scoring rules in the context of decision-making, as naive application of a proper scoring rule can lead to violation of first-order stochastic dominance. We argued that our results are relevant to experimental economists by offering an alternative interpretation of the data from Nyarko and Schotter (2002) and Palfrey and Wang (2009). Finally, we presented an estimator that may be used to identify the mean population belief from a group of risk averse forecasters when the population distribution of risk attitudes is known.

Possible lines of future research include adapting Proposition 4 to a setting with noisy reports and uncertainty about the risk aversion distribution. Additionally, we have not formally considered the relative merits of different scoring rules when risk aversion is present.

Scoring rules are important devices for applied research. The results in this paper should only help spur further investigations of their properties.

A Appendix

A.1 Proof of Lemma 1

We will make use of several facts from the proof of the more general Lemma 2. Note that Lagrange sufficiency (Assumption 5) is only needed to prove property *(iii)* in Proposition 2.

If we show that the univariate scoring rule $s(\cdot)$ is semi-strictly quasiconcave (Assumption 4) in its vector-valued form $\mathbf{s}(\cdot)$, the existence of the optimal report function as well as continuity and symmetry are immediate consequences of Proposition 2. Fix $p \in (0,1)$ (boundary cases are handled by Lemma 2). We see from relation (2) below that $s'(p) \neq 0$ for all $p \in (0,1)$. Since the scoring rule is continuously differentiable it must then either be the case that $s(\cdot)$ is everywhere strictly decreasing or everywhere strictly increasing. The former is clearly incompatible with strict propriety. Strict monotonicity implies strict quasiconcavity, and thus also that $\mathbf{s}(q, 1-q) = (s(q), s(1-q))$ is semi-strictly quasiconcave.

A.2 Proof of Proposition 1

Note first that the necessary first-order condition for $ps(q) + (1-p)s(1-q)$ to have an interior maximum at $q = p \in (0,1)$ is

$$\frac{p}{1-p} = \frac{s'(1-p)}{s'(p)}. \tag{2}$$

Property (i). Lemma 2 gives that $q(p) \in (0,1)$ if and only if $p \in (0,1)$. The necessary first-order condition for maximization of the expected utility $pu(s(q)) + (1-p)u(s(1-q))$ is

$$\frac{p}{1-p} = \frac{u'(s(1-q))}{u'(s(q))}\frac{s'(1-q)}{s'(q)} = \frac{u'(s(1-q))}{u'(s(q))}\frac{q}{1-q}. \tag{3}$$

The last equality uses (2). As $u'(\cdot)$ is decreasing and $s(\cdot)$ increasing, we see that $q(\cdot)$ must be increasing. Moreover, it is quickly verified that $q(1/2) = 1/2$ and $1/2 < q(p) < p$ whenever $p \in (1/2, 1)$.

Property (ii). If the utility function $\tilde{u}(\cdot)$ has a uniformly greater coefficient of absolute risk aversion than $u(\cdot)$ does, there exists an increasing and strictly concave function $\psi(\cdot)$ such that $\tilde{u}(\cdot) = \psi(u(\cdot))$ (Mas-Colell et al., 1995, Prop. 6.C.2). The condition (3) with $u = \tilde{u}$ is therefore

$$\frac{p}{1-p} = \frac{\psi'(u(s(1-q)))}{\psi'(u(s(q)))}\frac{u'(s(1-q))}{u'(s(q))}\frac{q}{1-q}.$$

Since $q_{\tilde{u}}(p) > 1/2$ for $p > 1/2$, and $\psi(\cdot)$ is strictly concave, the first factor on the right-hand side is greater than 1. The two last factors are increasing in q, so it must be the case that $q_{\tilde{u}}(p) < q_u(p)$ when $p > 1/2$.

A.3 Proof of Corollary 1

We shall use property *(iii)* of Proposition 1, namely that a uniformly higher coefficient of absolute risk aversion implies more shading.

Case (i). Assume that $u(\cdot)$ exhibits increasing relative risk aversion, i.e., that $-xu''(x)/u'(x)$ is increasing in x (the case of decreasing relative risk aversion is similar). Scaling the scoring rule by a constant $b > 0$ is mathematically equivalent to changing the utility function to $\tilde{u}_b(x) = u(bx)$ while keeping the scoring rule unchanged. Pick any $b, \hat{b} > 0$. Then $\tilde{u}_{\hat{b}}(\cdot)$ has a higher coefficient of absolute risk aversion than $\tilde{u}_b(\cdot)$ if

$$-\hat{b}\frac{u''(\hat{b}x)}{u'(\hat{b}x)} > -b\frac{u''(bx)}{u'(bx)}$$

at all x. By the assumption of increasing relative risk aversion of $u(\cdot)$, the above will hold whenever $\hat{b} > b$. If $u(\cdot)$ is of the constant relative risk aversion family, we can assume without loss of generality that $u(x) = x^{1-\gamma}/(1-\gamma)$ for some $\gamma > 0$ (with $\gamma = 1$ meaning logarithmic utility). We see that if the scoring rule is scaled by a constant $b > 0$, this constant will appear as a multiplicative factor $b^{1-\gamma}$ in the forecaster's objective function. It follows that the optimal choice of q will be independent of b.

Case (ii). Assume that $u(\cdot)$ exhibits increasing absolute risk aversion, i.e., $-u''(x)/u'(x)$ is increasing in x (the case of decreasing absolute risk aversion is similar). Shifting the scoring rule by a constant $a \in \mathbb{R}$ is equivalent to changing the utility function to $\tilde{u}_a(x) = u(x + a)$ while keeping the scoring rule unchanged. Hence, for any $a, \hat{a} \in \mathbb{R}$, $\tilde{u}_{\hat{a}}(\cdot)$ will have a higher coefficient of absolute risk aversion than $\tilde{u}_a(\cdot)$ if and only if $\hat{a} > a$. If $u(\cdot)$ belongs to the constant absolute risk aversion family, we may assume that $u(x) = -\exp(-\gamma x)$ for some $\gamma > 0$. If the scoring rule is shifted by a constant a, this constant will appear as a multiplicative factor $\exp(-\gamma a)$ in the forecaster's objective function, and so the optimal choice of q is invariant under shifts of $s(\cdot)$.

A.4 Proof of Lemma 2

We make the following observations. The differentiability assumptions ensure that $S_{\mathbf{p}}(\cdot)$ is continuously differentiable, and semi-strict quasiconcavity

of $\mathbf{s}(\cdot)$ and increasingness of $u(\cdot)$ imply that $S_{\mathbf{p}}(\cdot)$ is strictly quasiconcave (given interiority of $\mathbf{p}$). If a solution to the problem (4) exists, strict quasiconcavity of $S_{\mathbf{p}}(\cdot)$ and convexity of the constraint set guarantees that the maximizer is unique (Mas-Colell et al., 1995, Theorem M.K.4).

Before proceeding, we prove an auxiliary lemma.

Lemma A.1. *Let the scoring rule satisfy the assumptions of Lemma 2. If* $\mathbf{p} \in \Delta_n^\circ$*, then* $\mathbf{q} = \operatorname{argmax}_{\tilde{q} \in D} S_{\mathbf{p}}(\tilde{q})$ *exists and lies in* Δ_n°*.*

Proof. Observe first that since the scoring rule is extended continuous, we can extend the domain of a scoring rule defined on $D = \Delta_n^\circ$ to all of Δ_n, with the result being a scoring rule that is defined on a compact set. The exception is if $s_i(\mathbf{q}^k) \to -\infty$ for some $i = 1, \ldots, n$ and some sequence $\{\mathbf{q}^k\}$ tending to a boundary point.[17] However, in this case we can restrict the scoring rule to a closed subset of Δ_n on which we know for certain that $S_{\mathbf{p}}(\mathbf{q}) < S_{\mathbf{p}}(\mathbf{p})$, since the unboundedly negative scoring rule payoff near the problematic boundary points implies that there is some open neighborhood around each point in which the expected utility is very low.

Consequently, all reports that yield higher expected utility than truthful reporting lie in some compact set on which $S_{\mathbf{p}}(\cdot)$ may be defined, retaining continuity of the latter. Hence, an optimal report $\mathbf{q} \in \Delta_n$ exists.

Finally, we argue that $\mathbf{q} \in \Delta_n^\circ$. In the proof of Proposition 2 we show that an interior optimum must satisfy $q_i \geq \min_{j=1,\ldots,n} p_j$ for all $i = 1, \ldots, n$—cf. equation (6). Suppose that the optimal report has $q_l = 0$ for some $l = 1, \ldots, n$. Now consider a modified utility maximization problem in which the utility function is substituted with $\tilde{u}(x) = \theta x + (1-\theta)u(x)$, where $\theta \in [0, 1]$. For $\theta = 1$, strict propriety of the scoring rule implies that the optimal report is $\mathbf{q} = \mathbf{p}$. For $\theta = 0$ we have supposed that the optimal report is on the boundary. However, since the objective function is continuous in θ, the Theorem of the Maximum states that the optimal report must be continuous in θ, which would then imply $q_l < \min_{j=1,\ldots,n} p_j$ for sufficiently small θ, a contradiction. It follows that $q_i > 0$ for all $i = 1, \ldots, n$ in the original utility maximization problem. □

Given the above lemma, we first restrict attention to the case $\mathbf{p}, \mathbf{q} \in D = \Delta_n^\circ$ (except where noted), subsequently considering boundary reports.

Existence of optimal report function. This follows directly from Lemma A.1 when $\mathbf{p} \in \Delta_n^\circ$.

Continuity. The objective function in the problem (4) is continuous. According to the argument in the proof of Lemma A.1, for any $\mathbf{p} \in \Delta_n^\circ$

[17] The limit cannot be $+\infty$ since the scoring rule is bounded above.

we can construct a closed subset of Δ_n on which the optimal report lies, and on which we can continuously extend $S_{\mathbf{p}}(\cdot)$. This subset can be made continuous in $\mathbf{p}$. The Theorem of the Maximum (Mas-Colell et al., 1995, Theorem M.K.6) then states that $\mathbf{q}(\cdot)$ is continuous. Single-valuedness—and thus continuity—follows from uniqueness of the optimal report.

Symmetry. Using neutrality of the scoring rule, symmetry readily follows by inspection of the optimization problem (4).

Boundary reports. Now consider the case $D = \Delta_n$. Let $I = \{i = 1, \ldots, n \mid p_i = 0\}$. Suppose first that I is non-empty. We shall employ the notation

$$\Delta_n(\varepsilon) = \{\tilde{\mathbf{q}} \in \Delta_n \mid \tilde{q}_i \geq \varepsilon \text{ for all } i\}.$$

Note that $\Delta_n(\varepsilon) = \Delta_n$ for $\varepsilon \leq 0$. Define the correspondence $\mathbf{r}\colon \Delta_n \rightrightarrows \Delta_n$ by

$$\mathbf{r}(\tilde{\mathbf{p}}) = \operatorname*{argmax}_{\tilde{\mathbf{q}} \in \Delta_n(\min_k p_k)} S_{\tilde{\mathbf{p}}}(\tilde{\mathbf{q}})$$

for $\tilde{\mathbf{p}} \in \Delta_n$. Since the objective function and the constraint correspondence are continuous in $\tilde{\mathbf{p}}$, the Theorem of the Maximum states that $\mathbf{r}(\cdot)$ is upper hemicontinuous. Consider an interior sequence $\{\mathbf{p}^n\}$ converging to $\mathbf{p}$ as $n \to \infty$. Upper hemicontinuity implies that

$$\lim_{n\to\infty} \mathbf{r}(\mathbf{p}^n) \subset \mathbf{r}\left(\lim_{n\to\infty} \mathbf{p}^n\right) = \operatorname*{argmax}_{\tilde{\mathbf{q}} \in \Delta_n} S_{\mathbf{p}}(\tilde{\mathbf{q}}). \qquad (\dagger)$$

It is readily verified from relation (6) that, for each n, the unique unrestricted maximizer $\mathbf{q}^n$ of $S_{\mathbf{p}^n}(\tilde{\mathbf{q}})$ over Δ_n° satisfies $q_i^n \geq \min_k p_k$. Hence, $\mathbf{r}(\mathbf{p}^n) = \mathbf{q}^n$ (i.e., the correspondence is single-valued) and by expression (6) the i-th coordinate of that single value is proportional to p_i^n. Since $p_i^n \to 0$ for all $i \in I$, $\lim_{n\to\infty} \mathbf{r}(\mathbf{p}^n)$ consists of an element $\mathbf{q}^*$ with coordinates in I equal to zero. Suppose that there also existed a $\hat{\mathbf{q}} \in \operatorname{argmax}_{\tilde{\mathbf{q}}\in\Delta_n} S_{\mathbf{p}}(\tilde{\mathbf{q}})$ with $\hat{q}_i > 0$ for some $i \in I$. Then it must be the case that $q_j^* \neq \hat{q}_j$ for some $j \notin I$. Semi-strict quasiconcavity of the scoring rule then implies that $S_{\mathbf{p}}(\frac{1}{2}\mathbf{q}^* + \frac{1}{2}\hat{\mathbf{q}}) > S_{\mathbf{p}}(\mathbf{q}^*)$, a contradiction. Hence, the optimal report is unique and we have $q_i(\mathbf{p}) = 0$ for all $i \in I$.

Going the other way, suppose there existed a belief $\mathbf{p} \in \Delta_n$ such that some optimal report $\mathbf{q}$ had $q_i = 0$ even though $p_i > 0$. In light of the discussion above, we can without loss of generality consider the case $\mathbf{p} \in \Delta_n^\circ$. Thus, the objective function in (2) will be strictly quasiconcave on all of Δ_n, guaranteeing uniqueness of the optimal report. Define the correspondence $\tilde{\mathbf{r}}\colon \Delta_n \rightrightarrows \Delta_n$ by

$$\tilde{\mathbf{r}}(\tilde{\mathbf{p}}) = \operatorname*{argmax}_{\tilde{\mathbf{q}} \in \Delta_n(\tilde{p}_i - p_i)} S_{\tilde{\mathbf{p}}}(\tilde{\mathbf{q}})$$

for $\tilde{\mathbf{p}} \in \Delta_n$. Consider a sequence $\{\mathbf{p}^n\}$ converging to $\mathbf{p}$, with $p_i^n \in (p_i, 2p_i)$ at all n. Again, the Theorem of the Maximum gives that $\tilde{\mathbf{r}}(\cdot)$ is upper hemicontinuous, so (†) holds with $\mathbf{r}(\cdot)$ replaced with $\tilde{\mathbf{r}}(\cdot)$. But for each n, $\tilde{\mathbf{r}}(\mathbf{p}^n)$ is single-valued and the i-th coordinate of that value is proportional to p_i^n. It follows that there exists some $\mathbf{q}^* \in \operatorname{argmax}_{\tilde{\mathbf{q}} \in \Delta_n} S_{\mathbf{p}}(\tilde{\mathbf{q}})$ with $q_i^* > 0$. This contradicts uniqueness of $\mathbf{q}$.

In conclusion, we have shown that $q_i(\mathbf{p}) = 0$ if and only if $p_i = 0$. An immediate corollary is that $q_i(\mathbf{p}) = 1$ if and only if $p_i = 1$.

A.5 Proof of Proposition 2

Due to Lemma 2, we may restrict attention to interior subjective beliefs and reports. An expected utility maximizing forecaster solves

$$\max_{\tilde{\mathbf{q}} \in \mathbb{R}^n} S_{\mathbf{p}}(\tilde{\mathbf{q}}) \quad \text{s.t.} \quad \sum_{i=1}^{n} \tilde{q}_i = 1, \tilde{q}_i > 0 \text{ for all } i. \tag{4}$$

We start out by showing an auxiliary result.

Lemma 3. *Suppose the scoring rule satisfies the assumptions in Lemma 2. Let $\mathbf{q} \in \Delta_n^\circ$ and $i, j = 1, \ldots, n$. Then $s_i(\mathbf{q}) > s_j(\mathbf{q})$ if and only if $q_i > q_j$.*

Proof. If $q_i = q_j$, neutrality yields $s_i(\mathbf{q}) = s_j(\mathbf{q})$. When $q_i \neq q_j$, strict propriety of the scoring rule implies

$$\sum_{k=1}^{n} q_k s_k(\mathbf{q}) > \sum_{k=1}^{n} q_k s_k(\sigma_{ij}(\mathbf{q})).$$

Due to neutrality of the scoring rule, the $n-2$ terms corresponding to $k \neq i, j$ cancel out on both sides. Moreover, $s_i(\sigma_{ij}(\mathbf{q})) = s_j(\mathbf{q})$, and vice versa with i and j reversed. This leaves us with

$$(q_i - q_j)[s_i(\mathbf{q}) - s_j(\mathbf{q})] > 0,$$

from which the lemma follows. □

Let $\hat{\mathbf{q}} \in \Delta_n^\circ$. Since the scoring rule is Lagrange sufficient (and thus continuously differentiable), the necessary and sufficient first-order conditions for the problem (1) imply that if there exists a $\tilde{\mathbf{q}} \in \Delta_n^\circ$ and a multiplier $\mu \in \mathbb{R}$ such that, for all $l = 1, \ldots, n$,

$$\sum_{k=1}^{n} \hat{q}_k \frac{\partial s_k(\tilde{\mathbf{q}})}{\partial q_l} = \mu, \tag{5}$$

then it must be the case that $\tilde{\mathbf{q}} = \hat{\mathbf{q}}$.

Fix $\mathbf{p} \in \Delta_n^\circ$. As $S_{\mathbf{p}}(\cdot)$ is continuously differentiable, there must exist a $\lambda \in \mathbb{R}$ such that the unique solution $\mathbf{q} \in \Delta_n^\circ$ of the problem (4) satisfies the first-order conditions

$$\sum_{k=1}^{n} p_k u'(s_k(\mathbf{q})) \frac{\partial s_k(\mathbf{q})}{\partial q_l} = \lambda$$

for all $l = 1, \ldots, n$. Define $\hat{\mathbf{q}} \in \Delta_n^\circ$ by $\hat{q}_l = p_l u'(s_l(\mathbf{q})) / \sum_{k=1}^{n} p_k u'(s_k(\mathbf{q}))$, $l = 1, \ldots, n$, and divide the above conditions by $\sum_{k=1}^{n} p_k u'(s_k(\mathbf{q}))$ to find

$$\sum_{k=1}^{n} \hat{q}_k \frac{\partial s_k(\mathbf{q})}{\partial q_l} = \frac{\lambda}{\sum_{k=1}^{n} p_k u'(s_k(\mathbf{q}))}.$$

Since the right-hand side is independent of l, it follows from the previous discussion—cf. (5)—that $\hat{\mathbf{q}} = \mathbf{q}$, i.e.,

$$q_l = \frac{p_l u'(s_l(\mathbf{q}))}{\sum_{k=1}^{n} p_k u'(s_k(\mathbf{q}))} \tag{6}$$

for all $l = 1, \ldots, n$. Hence,

$$\frac{p_i}{p_j} = \frac{q_i}{q_j} \frac{u'(s_j(\mathbf{q}))}{u'(s_i(\mathbf{q}))}. \tag{7}$$

Since $u(\cdot)$ is strictly concave, $u'(\cdot)$ is strictly decreasing. Using Lemma 3, the equation readily implies that if $p_i > p_j$, then $p_i/p_j > q_i/q_j > 1$.

A.6 Proof of Proposition 3

We can without loss of generality focus on the case $n = 3$. For expositional simplicity, we also specify that the domain of the scoring rule is $D = \Delta_n$. Fix $\Omega = \{1, 2, 3\}$ and normalize the center's utility function $v(\cdot)$ so that $v(0) = 0$. The lottery $\tilde{L}$ is induced by probability distribution $\tilde{\mathbf{p}} = (1/3, 1/3, 1/3)$ and the act that pays out $\mathbf{x} = (x, x, 0)$, where $x > 0$. Lottery L is induced by $\mathbf{p} = (2/3 + \varepsilon, 0, 1/3 - \varepsilon)$ for some small $\varepsilon > 0$ and the same act. The latter lottery strictly first-order stochastically dominates the former.

From Propositions 1 and 2 it follows that $q_3(\tilde{\mathbf{p}}) = 1/3 < q_3(2/3, 0, 1/3)$, and so by continuity of $\mathbf{q}(\cdot)$ we have $q_3(\tilde{\mathbf{p}}) < q_3(\mathbf{p})$ for a sufficiently small ε. This implies

$$q_1(\mathbf{p})v(x) + q_2(\mathbf{p})v(x) + q_3(\mathbf{p}) \cdot 0 < q_1(\tilde{\mathbf{p}})v(x) + q_2(\tilde{\mathbf{p}})v(x) + q_3(\tilde{\mathbf{p}}) \cdot 0.$$

In other words, when deciding on the basis of $\mathbf{q}(\mathbf{p})$ and $\mathbf{q}(\tilde{\mathbf{p}})$, the center ranks lottery $\tilde{L}$ over L.

A.7 Proof of Proposition 4

The integral in the expression for the estimator is well-defined due to $q_\gamma^{-1}(\cdot)$ being continuous. Note first that[18]

$$q_\gamma^{-1}(\tilde{q}_1) = E[p_1 \mid \tilde{q}_1, \gamma_1 = \gamma].$$

Since the data points are i.i.d. and $|q_\gamma^{-1}(\cdot)|$ is bounded by 1 for all γ, the strong law of large numbers gives, as $N \to \infty$,

$$\begin{aligned}\hat{p} \overset{a.s.}{\to} E_{\tilde{q}_1}\left\{\int_\Gamma q_\gamma^{-1}(\tilde{q}_1)dG(\gamma)\right\} &= E_{\tilde{q}_1}\Big\{E_\gamma\big(E[p_1 \mid \tilde{q}_1, \gamma_1 = \gamma]\big)\Big\} \\ &= E_{\tilde{q}_1}\Big\{E[p_1 \mid \tilde{q}_1]\Big\} = E[p_1].\end{aligned}$$

A.8 Monte Carlo simulations of the Proposition 4 estimator

Monte Carlo simulations suggest that the estimator for the mean has good finite-sample properties, cf. Figure 2. In applications we expect an additional source of error stemming from uncertainty about the true distribution of risk attitudes. Proposition 1 provides guidance as to the direction of a possible bias. Figure 3 shows Monte Carlo results for a misspecified model where the degree of forecaster risk aversion has been overestimated by the center. Although consistency of the $\hat{p}$ estimator fails, the figure suggests that the error is quite manageable even in small samples.

[18] Given the assumption that $\tilde{q}_i = q_{\gamma_i}(p_i)$ exactly, the distribution of p_1 conditional on $\tilde{q}_1$ and γ_1 is degenerate. Hence, we have $[q_\gamma^{-1}(\tilde{q}_1)]^k = E[(p_1)^k \mid \tilde{q}_1, \gamma_1 = \gamma]$ for all k, a fact that may be used for deriving estimators of other moments of the belief distribution.

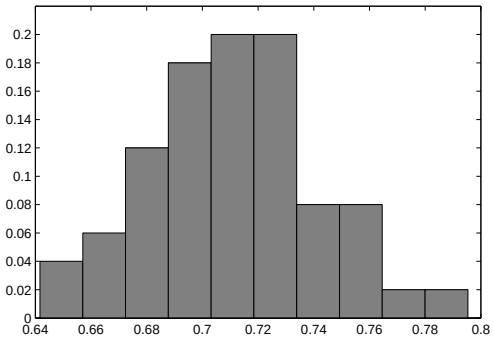

Figure 2: Histogram of the estimated population means in 50 runs of the Proposition 4 estimator on data with specifications $N = 20$, $u_\gamma(x) = -x^{-\gamma}$, $s(q) = 2 + 2q - q^2 - (1-q)^2$, $p_i \sim B(5,2)$, $\gamma_i \sim \log N(0.8, 0.4)$. The true population mean is $E[p_1] = 5/7 \approx 0.7143$. The average and standard deviation of the 50 estimates are 0.7108 and 0.0311, respectively. Calculations were carried out in Matlab using `quad` for numerical integration. For each γ and $\tilde{q}$ the inverse function $g_\gamma^{-1}(\tilde{q})$ was approximated by finding optimal reports at $p = 0.01, 0.02, \ldots, 1.00$ using `fminbnd` and then choosing the p that led to the smallest error $|q_\gamma(p) - \tilde{q}|$.

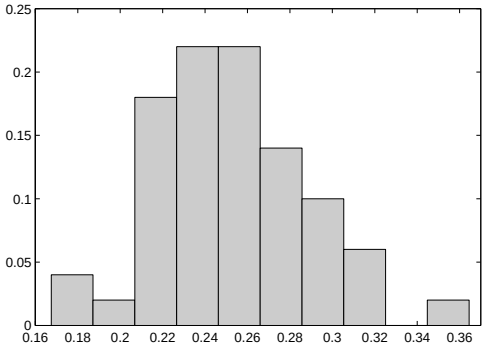

Figure 3: Histogram of the estimated population means in a *misspecified* model. The estimates derive from 50 runs of the Proposition 4 estimator on data with specifications $N = 20$, $u_\gamma(x) = -\exp(-\gamma x)$, $s(q) = q/\sqrt{q^2 + (1-q)^2}$, $p_i \sim B(1.5, 4)$, $\gamma_i \sim \Gamma(2, 1.5)$. However, the integration in the estimator is carried out with respect to the erroneous distribution $\gamma \sim \Gamma(2,2) = \chi^2(4)$. The true population mean is $E[p_1] = 1.5/5.5 \approx 0.2727$. The average and standard deviation of the 50 estimates are 0.2527 and 0.0380, respectively.

B References

References

Bickel, J. E. (2007), 'Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules', *Decision Analysis* **4**(2), 49–65.

Brier, G. W. (1950), 'Verification of forecasts expressed in terms of probability', *Monthly Weather Review* **78**(1), 1–3.

Costa-Gomes, M. A. and Weizsäcker, G. (2008), 'Stated Beliefs and Play in Normal-Form Games', *Review of Economic Studies* **75**(3), 729–762.

Gneiting, T. and Raftery, A. E. (2007), 'Strictly Proper Scoring Rules, Prediction, and Estimation', *Journal of the American Statistical Association* **102**, 359–378.

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

Hanson, R. (2007), 'Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation', *Journal of Prediction Markets* **1**(1), 1–15.

Holt, C. A. and Laury, S. K. (2002), 'Risk Aversion and Incentive Effects', *The American Economic Review* **92**(5), 1644–1655.

Holt, C. A. and Laury, S. K. (2005), 'Risk Aversion and Incentive Effects: New Data without Order Effects', *The American Economic Review* **95**(3), 902–904.

Jackwerth, J. C. (2000), 'Recovering Risk Aversion from Option Prices and Realized Returns', *The Review of Financial Studies* **13**(2), 433–451.

Kadane, J. B. and Winkler, R. L. (1988), 'Separating Probability Elicitation From Utilities', *Journal of the American Statistical Association* **83**(402), 357–363.

Kagel, J. H. and Roth, A. E., eds (1997), *The Handbook of Experimental Economics*, Princeton University Press.

Lerner, J. S. and Keltner, D. (2001), 'Fear, Anger and Risk', *Journal of Personality and Social Psychology* **81**(1), 146–159.

Manski, C. F. (2004), 'Measuring Expectations', *Econometrica* **72**(5), 1329–1376.

Mas-Colell, A., Whinston, M. D. and Green, J. R. (1995), *Microeconomic Theory*, Oxford University Press.

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

McKelvey, R. D. and Page, T. (1990), 'Public and Private Information: An Experimental Study of Information Pooling', *Econometrica* **58**(6), 1321–1339.

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

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

Offerman, T., Sonnemans, J., Kuilen, G. V. D. and Wakker, P. P. (2009), 'A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes', *Review of Economic Studies* **76**(4), 1461–1489.

Offerman, T., Sonnemans, J. and Schram, A. (1996), 'Value Orientations, Expectations and Voluntary Contributions in Public Goods', *Economic Journal* **106**(437), 817–45.

Palfrey, T. R. and Wang, S. W. (2009), 'On eliciting beliefs in strategic games', *Journal of Economic Behavior & Organization* **71**(2), 98–109.

Pennock, D. (2006), 'Implementing Hanson's Market Maker'. Oddhead Blog, October 30.
`blog.oddhead.com/2006/10/30/implementing-hansons-market-maker/`.

Prelec, D. (2004), 'A Bayesian Truth Serum for Subjective Data', *Science* **306**(5695), 462–466.

Roth, A. E. and Malouf, M. W. K. (1979), 'Game-Theoretic Models and the Role of Information in Bargaining', *Psychological Review* **86**(6), 574–594.

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

Savage, L. J. (1972), *The Foundations of Statistics*, Dover Publications, Inc.

Selten, R. (1998), 'Axiomatic Characterization of the Quadratic Scoring Rule', *Experimental Economics* **1**(1), 43–61.

Selten, R., Sadrieh, A. and Abbink, K. (1999), 'Money Does Not Induce Risk Neutral Behavior, but Binary Lotteries Do even Worse', *Theory and Decision* **46**, 213–252.

Smith, C. A. B. (1961), 'Consistency in Statistical Inference and Decision', *Journal of the Royal Statistical Society, Series B (Methodological)* **23**(1), 1–37.