Judgment and Decision Making, Vol. 10, No. 1, January 2015, pp. 86–96

Tailored proper scoring rules elicit decision weights

Arthur Carvalho*

Abstract

Proper scoring rules are scoring methods that incentivize honest reporting of subjective probabilities, where an agent strictly maximizes his expected score by reporting his true belief. The implicit assumption behind proper scoring rules is that agents are risk neutral. Such an assumption is often unrealistic when agents are human beings. Modern theories of choice under uncertainty based on rank-dependent utilities assert that human beings weight nonlinear utilities using decision weights, which are differences between weighting functions applied to cumulative probabilities.

In this paper, I investigate the reporting behavior of an agent with a rank-dependent utility when he is rewarded using a proper scoring rule tailored to his utility function. I show that such an agent misreports his true belief by reporting a vector of decision weights. My findings thus highlight the risk of utilizing proper scoring rules without prior knowledge about all the components that drive an agent's attitude towards uncertainty. On the positive side, I discuss how tailored proper scoring rules can effectively elicit weighting functions. Moreover, I show how to obtain an agent's true belief from his misreported belief once the weighting functions are known.

Keywords: proper scoring rules, rank-dependent utility theory, weighting functions.

1 Introduction

An agent's assessment of the likelihood of a future event in which he has no stake may be of interest to others. For example, a financial investor may be interested in the probability a market expert assigns to the increase of a certain stock price. In the medical domain, a patient might want to know the likelihood of success of a treatment before deciding whether to undergo that treatment.

Strategic agents are not necessarily honest when reporting their beliefs. For example, Nakazono (2013) reported that governors of the Federal Open Market Committee tend to report forecasts close to the previous consensus, whereas non-governors tend to report forecasts far away from the previous consensus. Nakazono concluded that both governors and non-governors behave strategically.

In cases where agents behave strategically, a method to promote honest reporting is crucial. *Proper scoring rules* are traditional scoring methods that induce honest reporting of subjective probabilities, in a sense that an agent maximizes his expected score from a proper scoring rule by reporting his true belief (Winkler & Murphy, 1968). Hence, the implicit assumption behind proper scoring rules is that agents are *risk neutral*, i.e., that they behave so as to maximize their expected scores.

The assumption of risk-neutral behavior is hardly compelling when the underlying agents are human beings. Several violations of risk neutrality have been reported in the literature (Allais, 1953; Holt & Laury, 2002; Starmer, 2000; Tversky & Kahneman, 1992). Winkler (1969) suggested an approach to tailor a proper scoring rule to an agent's nonlinear utility function. Under Winkler's approach, however, agents' utilities are still weighted by their subjective probabilities.

As I elaborate later in this paper, reporting a belief under a proper scoring rule is equivalent to making a choice under uncertainty. Consequently, one can analyze an agent's reporting behavior under different decision theories. Modern models of individual choices under uncertainty based on *rank-dependent utilities* assert that nonlinear utility functions are weighted by *decision weights*, instead of subjective probabilities (Quiggin, 1982; Schmeidler, 1989). Decision weights are differences between *weighting functions* applied to cumulative probabilities. Thus, according to traditional rank-dependent models, an agent's attitude towards uncertainty is driven by both a utility function and weighting functions.

In this paper, I investigate how an agent who makes decisions based on a rank-dependent utility reports his belief under a proper scoring rule tailored to his utility function. I show that such an agent misreports his true belief by reporting a vector of decision weights. Decision weights reflect a cognitive bias concerning how human beings deal with probabilities when making choices under uncertainty and, thus, they should not be taken as a measure of an agent's true belief. Thus, my findings highlight the neces-

The author acknowledges helpful comments and suggestions from Jonathan Baron, Michael H. Birnbaum, and an anonymous reviewer.

*Rotterdam School of Management, Erasmus University, 3062 PA, Rotterdam, The Netherlands. Email: carvalho@rsm.nl.

sity of knowing all the components that drive an agent's attitude towards uncertainty before appropriately using a proper scoring rule to elicit that agent's belief.

On the positive side, I show how a proper scoring rule tailored to an agent's utility function can effectively elicit that agent's weighting functions. Moreover, I suggest recursive procedures to obtain the agent's true belief once his weighting functions are known.

2 Related work

The task of inducing honest reporting of private information has been extensively studied in the fields of mechanism design and decision theory. My focus in this paper is on the elicitation of private information as subjective probabilities (beliefs) over uncertain outcomes.

Proper scoring rules provide a prominent technique to induce honest reporting of subjective probabilities. Proper scoring rules have been used in a variety of domains, e.g., when sharing rewards amongst a set of agents based on peer evaluations (Carvalho & Larson, 2010, 2011, 2012), when incentivizing agents to accurately estimate their own efforts to accomplish a task (Bacon et al., 2012), to elicit opinions from policy makers regarding the occurrence of political and economic events (Tetlock, 2005), etc.

A standard assumption when using proper scoring rules is that agents are risk neutral. Focusing on the quadratic scoring rule, Winkler and Murphy (1970) investigated the effects of nonlinear utilities on how agents report their beliefs. More precisely, for some specific utility functions, Winkler and Murphy (1970) showed that a risk-seeking agent reports a very sharp probability distribution, whereas a risk-averse agent reports a probability distribution close to the uniform distribution. Winkler (1969) discussed how any proper scoring rule can be adjusted to an agent's nonlinear utility function, resulting in what I refer to in this paper as *tailored proper scoring rules*.

The aforementioned works are still within the expected utility theory framework. Modern theories of choice under uncertainty based on rank-dependent utilities assert that, aside from nonlinear utilities, probability sensitivity also plays a role in defining an agent's attitude towards uncertainty (Quiggin, 1982; Schmeidler, 1989). Focusing on binary outcomes, Offerman, Sonnemans, Van De Kuilen, and Wakker (2009) discussed how to calibrate *a posteriori* beliefs reported under the quadratic scoring rule by agents who take decisions based on rank-dependent utilities. Kothiyal, Spinu, and Wakker (2011) extended the work by Offerman et al. (2009) to all positive proper scoring rules. Moreover, Kothiyal et al. (2011) briefly mentioned that agents with rank-dependent utilities report vectors of decision weights instead of their true beliefs for the specific case when their utility functions are linear.

I generalize the results of Kothiyal et al. (2011) to any proper scoring rule, any finite number of outcomes, and any strictly increasing utility function. More specifically, I show that, when the utility function of an agent who makes decisions based on a rank-dependent utility is known and incorporated into a proper scoring rule, the agent still misreports his belief by reporting a vector of decision weights. Such reporting behavior happens because probability sensitivity, which is defined in terms of weighting functions, plays a crucial role when an agent reports his belief under a proper scoring rule.

I also show how to elicit weighting functions using tailored proper scoring rules. A popular method for eliciting weighting functions was proposed by Abdellaoui (2000). Abdellaoui's method implicitly assumes that agents are honest when reporting indifferences between lotteries. My approach, on the other hand, is based on the reports of beliefs for events with known objective probabilities (decision under risk), and honest reporting maximizes an agent's rank-dependent utility, thus resulting in a more reliable elicitation process.

3 Proper scoring rules

Consider a set of exhaustive and mutually exclusive outcomes $\theta_1, \theta_2, \ldots, \theta_n$, for $n \geq 2$. I assume that agents have *beliefs* (subjective probabilities) regarding the occurrence of the outcomes. Formally, an agent's belief is the probability vector $\mathbf{p} = (p_1, \ldots, p_n)$, where p_k is his subjective probability regarding the occurrence of outcome θ_k. Agents are self-interested and, consequently, they are not necessarily honest when reporting their beliefs. Therefore, I distinguish between an agent's *true belief* $\mathbf{p}$, and his *reported belief* $\mathbf{q} = (q_1, \ldots, q_n)$.

Proper scoring rules are traditional devices used to promote honest reporting of subjective probabilities (Winkler & Murphy, 1968). Formally, a *scoring rule* $R(\mathbf{q}, \theta_x)$ is a function that provides a *score* for the reported belief $\mathbf{q}$ upon observing the outcome θ_x. Scores are somehow coupled with relevant incentives, be they social-psychological, such as praise or visibility, or material rewards through prizes or money. A scoring rule is called *proper* when an agent maximizes his expected score (according to his own beliefs) by reporting a belief $\mathbf{q}$ that corresponds to his true belief $\mathbf{p}$ (Winkler & Murphy, 1968). A *strictly proper scoring rule* means that an agent maximizes his expected score if and only if he reports $\mathbf{q} = \mathbf{p}$. The expected score of an agent for a real-valued scoring rule $R(\mathbf{q}, \theta_x)$ is:

$$\mathbb{E}_{\mathbf{p}}\left[R(\mathbf{q}, \cdot)\right] = \sum_{k=1}^{n} p_k \, R(\mathbf{q}, \theta_k) \tag{1}$$

The best known strictly proper scoring rules, together with their scoring ranges, are:

$$\text{spherical: } R(\mathbf{q}, \theta_x) = \frac{q_x}{\sqrt{\sum_{k=1}^{n} q_k^2}} \qquad [0, 1]$$

$$\text{logarithmic: } R(\mathbf{q}, \theta_x) = \log q_x \qquad (-\infty, 0]$$

$$\text{quadratic: } R(\mathbf{q}, \theta_x) = 2q_x - \sum_{k=1}^{n} q_k^2 \qquad [-1, 1]$$

For the sake of illustration, consider a coin toss experiment with two outcomes ($n = 2$): $\theta_1 =$ "heads" and $\theta_2 =$ "tails". Consider that an agent i has a true belief $\mathbf{p} = (0.4, 0.6)$. Assume that agent i reports the belief $\mathbf{q} = (q_1, q_2)$, which is rewarded according to the logarithmic scoring rule. Then, agent i's expected score is $\mathbb{E}_{\mathbf{p}}[R(\mathbf{q}, \cdot)] = p_1 \log q_1 + p_2 \log q_2 = 0.4 \log q_1 + 0.6 \log q_2$. In the future, if outcome θ_1 is the observed outcome, then the score agent i receives is equal to $\log q_1$. Since the logarithmic scoring rule is a strictly proper scoring rule, agent i's expected score is strictly maximized when he is honest, i.e., when $\mathbf{q} = \mathbf{p} = (0.4, 0.6)$. To show this, note that $\mathbb{E}_{\mathbf{p}}[R(\mathbf{q}, \cdot)] = p_1 \log q_1 + p_2 \log q_2 = p_1 \log q_1 + (1 - p_1) \log(1 - q_1)$. Since the resulting expected score is concave in q_1, the value of q_1 that maximizes agent i's expected score can be found by taking the first-order derivative of $\mathbb{E}_{\mathbf{p}}[R(\mathbf{q}, \cdot)]$ with respect to q_1, and equating the result to zero, i.e.:

$$\frac{p_1}{q_1} - \frac{1 - p_1}{1 - q_1} = 0 \implies p_1 = q_1$$

Selten (1998) and Jose (2009) provided axiomatic characterizations of, respectively, the quadratic scoring rule and the spherical scoring rule in terms of desirable properties, e.g., sensitivity to small probability values, symmetry, etc. In a seminal work, Savage (1971) showed that any differentiable strictly convex function $J(\mathbf{q})$ that is well-behaved at the endpoints of the scoring range can be used to generate a proper scoring rule. Formally:

$$R(\mathbf{q}, \theta_x) = J(\mathbf{q}) - \left(\sum_{k=1}^{n} \frac{\partial J(\mathbf{q})}{\partial q_k} \times q_k \right) + \frac{\partial J(\mathbf{q})}{\partial q_x}$$

For example, the logarithmic scoring rule can be derived from $J(\mathbf{q}) = \sum_{k=1}^{n} q_k \log q_k$:

$$\begin{aligned} R(\mathbf{q}, \theta_x) &= \sum_{k=1}^{n} q_k \log q_k - \left(\sum_{k=1}^{n} (\log q_k + 1) \times q_k \right) + \\ &\quad \log q_x + 1 \\ &= \log q_x \end{aligned}$$

I say that a scoring rule is *positive* when all the returned scores are nonnegative, i.e., $R(\mathbf{q}, \theta_x) \geq 0$ for all $x \in \{1, \ldots, n\}$. The spherical scoring rule is an example of a positive scoring rule. A *negative scoring rule*, on the other hand, returns only nonpositive scores, i.e., $R(\mathbf{q}, \theta_x) \leq 0$ for all $x \in \{1, \ldots, n\}$. The logarithmic scoring rule is an example of a negative scoring rule. Finally, a *mixed scoring rule* might return both positive and negative scores. The quadratic scoring rule is an example of a mixed scoring rule.

On a side note, I observe that proper scoring rules not only induce honest reporting of subjective probabilities, but they also measure the accuracy of reported beliefs, a task often called forecast verification. In particular, the more an agent moves probability mass to the observed outcome, the higher the agent's score will be.

3.1 Tailored proper scoring rules

An implicit assumption in the definition of proper scoring rules is that agents are *risk neutral*, i.e., they report their beliefs so as to maximize their expected scores. Since $\arg\max_{\mathbf{q}} \mathbb{E}_{\mathbf{p}}[R(\mathbf{q}, \cdot)] = \mathbf{p}$, a risk-neutral agent has to honestly report his belief under a proper scoring rule R in order to maximize his expected score. Regarding risk neutrality, Savage (1971) said the following in his seminal work about the theoretical foundations of proper scoring rules:

> "This assumption is not altogether unobjectionable; for it may imply that the person's utility function is linear in money. But such linearity assumptions are made almost throughout the present paper and are presumably tolerable if only moderate sums of money are involved." (Savage, 1971, page 791)

In other words, the function that represents the value that an agent derives from a score, called the *utility function*, is linear with respect to the range of the score used in conjunction with the scoring rule. Theoretically, an agent's utility function is approximately linear when the stakes are low (Arrow, 1971, page 100). In practice, however, human beings' utility functions tend to become nonlinear when the stakes are high (Wakker, 2010, §2).

Expected utility theory tackles some of the problems concerning risk neutrality by assuming that utility functions might be nonlinear. More specifically, the curvature of the utility function determines an agent's attitude towards uncertainty, e.g., a convex utility function implies that the agent is *risk seeking*, whereas a concave utility function indicates that the agent is *risk averse*. Risk-neutral behavior arises only when the utility function is linear. Naturally, agents are assumed to behave so as to maximize their expected utilities.

In the context of proper scoring rules, an agent who behaves according to expected utility theory reports a belief $\mathbf{q}$ so that $\mathbf{q} = \arg\max_{\mathbf{z}} \mathbb{E}_{\mathbf{p}}[U(R(\mathbf{z}, \cdot))]$, where $U(\cdot)$ is the agent's utility function. Often in this setting, proper scoring rules are no longer proper, i.e., there are cases where $\arg\max_{\mathbf{z}} \mathbb{E}_{\mathbf{p}}\left[U(R(\mathbf{z},\cdot))\right] \neq \mathbf{p}$ (Winkler & Murphy, 1970). Winkler (1969) discussed how the composite function $S = U^{-1} \circ R$ is a proper scoring rule under a strictly increasing utility function U. That is, the scoring rule $S(\mathbf{q},\theta_x)$ is tailored to the agent's utility function. For example, consider the logarithmic scoring rule $R(\mathbf{q},\theta_x) = \log q_x$, and a concave utility function $U(y) = \log y$. Then, the *tailored proper scoring rule*[1] is:

$$S(\mathbf{q},\theta_x) = U^{-1}\left(R(\mathbf{q},\theta_x)\right) = e^{\log q_x} = q_x$$

Clearly, tailored proper scoring rules subsume traditional proper scoring rules since the latter assume that utility functions are linear. In the following sections, I study the reporting behavior of agents under tailored proper scoring rules. Thus, an implicit assumption in my analysis is that an agent's utility function is known *a priori*, for example, it was previously elicited using an approach such as the tradeoff method (Wakker & Deneffe, 1996). However, I make no assumptions on U, except that it is a strictly increasing function, which implies that there exists an inverse function U^{-1} defined over the range of the utility function U.

4 Rank-dependent utility

When selecting and reporting a probability vector $\mathbf{q}$ under a tailored proper scoring rule, an agent is essentially taking a decision under uncertainty, where the potential payoffs resulting from his choice are defined by $S(\mathbf{q},\theta_x)$, for $x \in \{1,\ldots,n\}$. Consequently, an agent's reporting behavior can be analyzed from the perspective of different decision theories under uncertainty.

Unarguably, expected utility theory represents a crucial advancement in decision theory under uncertainty. Expected utility theory suggests an elegant and simple way of combining subjective probabilities and payoffs into a single measure of value, which has a number of appealing theoretical properties. However, several violations of the premises of expected utility theory have been widely reported. Many of these violations, such as the common consequence effect and the common ratio effect, can be explained by models that take subjective attitudes to probability into account, such as *rank-dependent models* (Quiggin, 1982; Schmeidler, 1989).

Rank-dependent models assert that both sensitivity to payoffs and sensitivity to probabilities generate deviations from risk neutrality. In particular, these models convert subjective probabilities into *decision weights*, and agents are assumed to take decisions so as to maximize their *rank-dependent utilities* (RDU). A possible interpretation of decision weights is that they represent a cognitive bias concerning how human beings deal with probability values when making choices under risk and uncertainty.

Rank-dependent models are amongst the most satisfactory decision theories under uncertainty (but, as discussed later, other models may be better still). Starmer (2000) and Camerer (2004) documented the superior predictive performance of rank-dependent models over expected utility theory for a range of phenomena, including the disposition effect, the equity premium puzzle, asymmetric price elasticities, the excess sensitivity of consumption to income, elasticities of labour supply and asset pricing, etc.

By construction, rank-dependent models can explain everything that expected utility theory can, but the converse is false. Under expected utility theory, an agent reports his true belief under a tailored proper scoring rule. In the next sections, I show that this is no longer the case under a rank-dependent model. In order to build intuition, I first introduce RDU in terms of *lotteries*, which are event-contingent payoffs. Thereafter, I extend the initial definition of RDU to tailored proper scoring rules and characterize how an underlying agent reports his belief.

4.1 RDU and lotteries

Let $\mathbf{l} = [y_1 : \theta_1, \ldots, y_n : \theta_n]$ denote a *lottery* which yields a payoff of $y_x \in \Re$ if outcome θ_x occurs. Since one can always rearrange the outcomes, I assume without loss of generality that $y_n \geq y_{n-1} \geq \cdots \geq y_1$. Given that agents have beliefs over the occurrence of the outcomes, I can then represent a lottery as $\mathbf{l} = [y_1 : p_1, \ldots, y_n : p_n]$, which yields a payoff of $y_x \in \Re$ with probability p_x.

A lottery is called *positive* when all payoffs are nonnegative, i.e., $y_n \geq y_{n-1} \geq \cdots \geq y_1 \geq 0$. I denote a positive lottery by $\mathbf{l}^+$. A lottery is called *negative* when all payoffs are nonpositive, i.e., $0 \geq y_n \geq y_{n-1} \geq \cdots \geq y_1$. I denote a negative lottery by $\mathbf{l}^-$. Finally, a *mixed lottery* $\mathbf{l}^\pm$ contains both positive and negative payoffs, i.e., $y_n \geq y_{n-1} \geq \cdots \geq y_i \geq 0 \geq y_{i-1} \geq \cdots \geq y_1$.

Focusing first on positive lotteries, rank-dependent models state that the value that a human being assigns to $\mathbf{l}^+$ is described according to his *rank-dependent utility* (RDU) (Quiggin, 1982):

$$RDU\left(\mathbf{l}^+\right) = \sum_{k=1}^{n} \pi_k^+ U(y_k) \qquad (2)$$

[1] The term tailored scoring rule was first used by Johnstone, Jose, and Winkler (2011) to describe a proper scoring rule tailored to a specific decision-making problem. My definition is different in that a tailored proper scoring rule is tailored to an agent's utility function. It is also noteworthy that my setting is different than the scenario described by Johnstone (2011), where an agent (forecaster) might have to consider an user's utility function when reporting his belief (forecast).

where:

$$\begin{aligned} \pi_n^+ &= W^+(p_n) \\ \pi_k^+ &= W^+\left(\sum_{x=k}^{n} p_x\right) - W^+\left(\sum_{x=k+1}^{n} p_x\right), \end{aligned} \tag{3}$$

for $k \in \{1, \ldots, n-1\}$. The function $W^+\colon [0,1] \to [0,1]$, also known as the *weighting function*, is striclty increasing, and it satisfies $W^+(0) = 0$ and $W^+(1) = 1$. Henceforth, I drop the superscript whenever talking about weighting functions in general, and not only in the domain of gains. As suggested by Gonzalez and Wu (1999), the weighting functions model the "psychophysics of chance", i.e., the way human beings subjectively distort probability values. Common findings suggest that the weighting function is a nonlinear transformation of the probability scale that overweights small probabilities and underweights moderate and high probabilities (Tversky & Kahneman, 1992; Abdellaoui, 2000). In other words, the weighting function displays an inverse-S shape: it is concave near 0 and convex near 1. The weighting function proposed by Tversky and Kahneman (1992) is:

$$W(\rho) = \frac{\rho^\gamma}{\left(\rho^\gamma + (1-\rho)^\gamma\right)^{\frac{1}{\gamma}}} \tag{4}$$

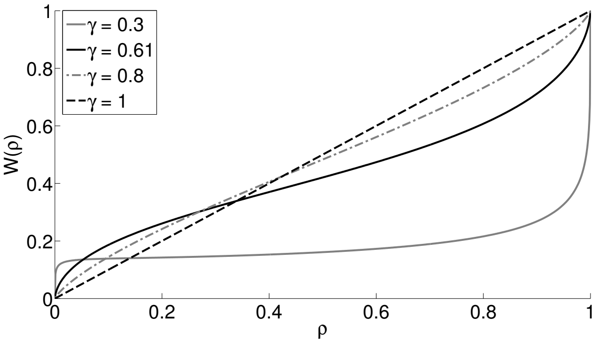
where $\gamma \geq 0.28$ in order for W to be strictly increasing. For $\gamma = 1$, the weighting function in (4) becomes the identity function. Decreasing γ results in a more pronounced inverse-S shape. Figure 1 illustrates the weighting function in (4) for different values of γ.

There are two crucial points regarding the rank-dependent utility in (2). First, as in the expected utility theory, the value that an agent derives from a payoff in a lottery is given by a strictly increasing utility function $U\colon \Re \to \Re$. Second, instead of an individual probability value p_k as in the expected utility theory, the weight of a utility $U(y_k)$ in (2) is the difference between two transformed *ranks*, $W(p_k + \cdots + p_n) - W(p_{k+1} + \cdots + p_n)$, also called a *decision weight*. For a lottery $\mathbf{l}^+$, the rank of a payoff y_k is the probability of $\mathbf{l}^+$ yielding a payoff better than y_k, i.e., the rank of y_k is equal to $p_{k+1} + p_{k+2} + \cdots + p_n$. The weight of $U(y_k)$ is then the transformed marginal contribution of the individual probability p_k to the total probability of receiving payoffs better than y_k.

Under rank-dependent models, positive and negative lotteries might be evaluated differently. For a negative lottery $\mathbf{l}^-$, the rank-depend utility in (2) is now defined as:

$$RDU\left(\mathbf{l}^-\right) = \sum_{k=1}^{n} \pi_k^- U(y_k)$$

Figure 1: The weighting function in (4) for different parameter values γ.



where:

$$\begin{aligned} \pi_1^- &= W^-(p_1) \\ \pi_k^- &= W^-\left(\sum_{x=1}^{k} p_x\right) - W^-\left(\sum_{x=1}^{k-1} p_x\right), \end{aligned} \tag{5}$$

for $k \in \{2, \ldots, n\}$. While a decision weight π_k^+ denotes the marginal contribution of an individual probability value p_k to the total probability of receiving better payoffs, a decision weight π_k^- denotes the marginal contribution of an individual probability value p_k to the total probability of receiving worse payoffs, measured in terms of a weighting function $W^-\colon [0,1] \to [0,1]$.

Finally, for a mixed lottery $\mathbf{l}^\pm$, where $y_n \geq y_{n-1} \geq \cdots \geq y_i \geq 0 \geq y_{i-1} \geq \cdots \geq y_1$, the rank-depend utility is now defined as:

$$RDU\left(\mathbf{l}^\pm\right) = \sum_{k=1}^{i-1} \pi_k^- U(y_k) + \sum_{k=i}^{n} \pi_k^+ U(y_k)$$

4.2 RDU and tailored proper scoring rules

Without loss of generality due to a possible rearrangement of outcomes, assume that the scores from a tailored proper scoring rule S are ordered, i.e., $S(\mathbf{q}, \theta_n) \geq S(\mathbf{q}, \theta_{n-1}) \geq \cdots \geq S(\mathbf{q}, \theta_1)$. I note that the scores from a tailored proper scoring rule can be stated in terms of a lottery: $[S(\mathbf{q}, \theta_1)\colon p_1, \ldots, S(\mathbf{q}, \theta_n)\colon p_n]$. Consequently, when reporting a belief $\mathbf{q}$, an agent is essentially defining the payoffs of a lottery, where the associated probabilities are subjective probabilities. In other words, reporting a belief $\mathbf{q}$ is equivalent to choosing a lottery amongst a potentially infinite number of lotteries. This implies that an agent's reporting behavior can be analyzed from the perspective of decision models such as rank-dependent models. For a positive, tailored proper scoring rule $S(\mathbf{q}, \theta_x)$, the rank-dependent utility in (2) becomes:

$$\sum_{k=1}^{n} \pi_k^+ U\left(S\left(\mathbf{q}, \theta_k\right)\right) \tag{6}$$

Similarly, the RDU for a negative, tailored proper scoring rule $S(\mathbf{q}, \theta_x)$ is:

$$\sum_{k=1}^{n} \pi_k^- U\left(S\left(\mathbf{q}, \theta_k\right)\right) \tag{7}$$

Finally, the RDU for a mixed, tailored proper scoring rule $S(\mathbf{q}, \theta_x)$ is:

$$\sum_{k=1}^{i-1} \pi_k^- U\left(S(\mathbf{q}, \theta_k)\right) + \sum_{k=i}^{n} \pi_k^+ U\left(S(\mathbf{q}, \theta_k)\right) \tag{8}$$

From the above equations, one might expect that an agent who maximizes a rank-dependent utility will behave differently than an expected-utility maximizer and, consequently, will report a belief other than his true belief under a tailored proper scoring rule. I discuss this point in the following section.

5 Characterizing reporting behavior under tailored proper scoring rules and RDU

The following propositions characterize how an agent who behaves to maximize a rank-dependent utility reports his belief under a tailored proper scoring rule. In short, my results indicate that such an agent reports a vector of decision weights, instead of his true belief.

Proposition 1. *Let $S(\mathbf{q}, \theta_x)$ be a positive, tailored proper scoring rule where $S(\mathbf{q}, \theta_n) \geq S(\mathbf{q}, \theta_{n-1}) \geq \cdots \geq S(\mathbf{q}, \theta_1) \geq 0$. Assume that an agent reports his belief $\mathbf{q}$ so as to maximize his RDU shown in (6). Then,*

$$\arg\max_{\mathbf{q}} \sum_{k=1}^{n} \pi_k^+ U\left(S(\mathbf{q}, \theta_k)\right) = (\pi_1^+, \pi_2^+, \ldots, \pi_n^+).$$

Proof. I start by noting that $U\left(S(\mathbf{q}, \theta_x)\right) = U\left(U^{-1}\left(R(\mathbf{q}, \theta_x)\right)\right) = R(\mathbf{q}, \theta_x)$, for some proper scoring rule R. If $\boldsymbol{\pi}^+ = (\pi_1^+, \pi_2^+, \ldots, \pi_n^+)$ is a probability vector, then $\sum_{k=1}^{n} \pi_k^+ R(\mathbf{q}, \theta_k) = \mathbb{E}_{\boldsymbol{\pi}^+}\left[R(\mathbf{q}, \cdot)\right]$, as in equation (1), and, consequently, $\arg\max_{\mathbf{q}} \sum_{k=1}^{n} \pi_k^+ R(\mathbf{q}, \theta_k) = \boldsymbol{\pi}^+$. Thus, I just need to prove that $\boldsymbol{\pi}^+ = (\pi_1^+, \ldots, \pi_n^+)$ is indeed a probability vector. From (3), I deduce that $\sum_{k=1}^{n} \pi_k^+ = W^+\left(\sum_{k=1}^{n} p_k\right) = 1$. Since W^+ is a strictly increasing function and its image is equal to $[0, 1]$, then $0 \leq \pi_k^+ \leq 1$, for all $k \in \{1, \ldots, n\}$, thus completing the proof. □

A similar result holds for negative, tailored proper scoring rules, as shown in Proposition 2.

Proposition 2. *Let $S(\mathbf{q}, \theta_x)$ be a negative, tailored proper scoring rule where $0 \geq S(\mathbf{q}, \theta_n) \geq S(\mathbf{q}, \theta_{n-1}) \geq \cdots \geq S(\mathbf{q}, \theta_1)$. Assume that an agent reports his belief $\mathbf{q}$ so as to maximize his RDU shown in (7). Then,*

$$\arg\max_{\mathbf{q}} \sum_{k=1}^{n} \pi_k^- U\left(S(\mathbf{q}, \theta_k)\right) = (\pi_1^-, \pi_2^-, \ldots, \pi_n^-).$$

Proof. Given that $U\left(S(\mathbf{q}, \theta_x)\right) = R(\mathbf{q}, \theta_x)$, for some proper scoring rule R, if $\boldsymbol{\pi}^- = (\pi_1^-, \pi_2^-, \ldots, \pi_n^-)$ is a probability vector, then $\sum_{k=1}^{n} \pi_k^- R(\mathbf{q}, \theta_k) = \mathbb{E}_{\boldsymbol{\pi}^-}\left[R(\mathbf{q}, \cdot)\right]$, as in equation (1). Consequently, $\arg\max_{\mathbf{q}} \sum_{k=1}^{n} \pi_k^- R(\mathbf{q}, \theta_k) = \boldsymbol{\pi}^-$. Thus, I just need to prove that $\boldsymbol{\pi}^- = (\pi_1^-, \ldots, \pi_n^-)$ is indeed a probability vector. From (5), I deduce that $\sum_{k=1}^{n} \pi_k^- = W^-\left(\sum_{k=1}^{n} p_k\right) = 1$. Since W^- is a strictly increasing function and its image is equal to $[0, 1]$, then $0 \leq \pi_k^- \leq 1$, for all $k \in \{1, \ldots, n\}$, thus completing the proof. □

Propositions 1 and 2 imply that positive and negative tailored proper scoring rules induce different reporting behavior whenever the weighting functions W^+ and W^- are different. In other words, a simple positive affine transformation of a proper scoring rule might induce different reporting behavior. I illustrate this point in Section 5.1. I show in the following proposition that mixed, tailored proper scoring rules induce agents to report decision weights as well when $W^+(\rho) + W^-(1 - \rho) = 1$, for all $\rho \in [0, 1]$.

Proposition 3. *Let $S(\mathbf{q}, \theta_x)$ be a mixed, tailored proper scoring rule where $S(\mathbf{q}, \theta_n) \geq S(\mathbf{q}, \theta_{n-1}) \geq \cdots \geq S(\mathbf{q}, \theta_i) \geq 0 \geq S(\mathbf{q}, \theta_{i-1}) \geq \cdots \geq S(\mathbf{q}, \theta_1)$. Assume that an agent reports his belief $\mathbf{q}$ so as to maximize his RDU shown in (8). If $W^+(\rho) + W^-(1 - \rho) = 1$, for any $\rho \in [0, 1]$, then*

$$\arg\max_{\mathbf{q}} \left(\sum_{k=1}^{i-1} \pi_k^- U\left(S(\mathbf{q}, \theta_k)\right) + \sum_{k=i}^{n} \pi_k^+ U\left(S(\mathbf{q}, \theta_k)\right) \right)$$
$$= (\pi_1^-, \ldots, \pi_{i-1}^-, \pi_i^+, \ldots, \pi_n^+).$$

Proof. If $\boldsymbol{\pi}^\pm = \left(\pi_1^-, \ldots, \pi_{i-1}^-, \pi_i^+, \ldots, \pi_n^+\right)$ is a probability vector, then the result follows naturally because $U\left(S(\mathbf{q}, \theta_x)\right) = R(\mathbf{q}, \theta_x)$, for some proper scoring rule R. Consequently, I just need to prove that $\boldsymbol{\pi}^\pm$ is indeed a probability vector. From (3) and (5), I have that $\sum_{k=1}^{i-1} \pi_k^- + \sum_{k=i}^{n} \pi_k^+ = W^-\left(\sum_{k=1}^{i-1} p_k\right) + W^+\left(\sum_{k=i}^{n} p_k\right) = 1$, where the last equality follows from the assumption that $W^+(\rho) + W^-(1 - \rho) = 1$, for all $\rho \in [0, 1]$. Since both W^+ and W^- are strictly increasing functions and their images are equal to $[0, 1]$, then $0 \leq \pi_j^-, \pi_k^+ \leq 1$, for all $j \in \{1, \ldots, i-1\}$ and $k \in \{i, \ldots, n\}$, thus completing the proof. □

5.1 Numerical example

In this subsection, I illustrate the theoretical results proved in Propositions 1 and 2 by using the weighting function proposed by Tversky and Kahneman (1992) shown in (4). Tversky and Kahneman (1992) found that the best fit for their data happened when using W^+ and W^- as defined in (4) with parameter values equal to, respectively, $\gamma = 0.61$ and $\gamma = 0.69$.

Consider an agent with belief $\mathbf{p} = (0.2, 0.8)$ who behaves so as to maximize his rank-dependent utility. Under a positive, tailored proper scoring rule, Proposition 1 implies that the agent reports:

$$\mathbf{q} = (1 - W^+(0.8), W^+(0.8)) = (0.393, 0.607)$$

Proposition 2 implies that the same agent reports:

$$\mathbf{q} = (W^-(0.2), 1 - W^-(0.2)) = (0.257, 0.743)$$

under a negative, tailored proper scoring rule. The deviation of the agent's reported belief $\mathbf{q}$ from his true belief $\mathbf{p}$ according to the mean absolute error is equal to: $0.5 \times |1 - W^+(0.8) - 0.2| + 0.5 \times |W^+(0.8) - 0.8| = 0.193$, for a positive, tailored proper scoring rule, and $0.5 \times |W^-(0.2) - 0.2| + 0.5 \times |1 - W^-(0.2) - 0.8| = 0.057$ for a negative, tailored proper scoring rule.

The above example illustrates that positive and negative tailored proper scoring rules might induce different reporting behavior whenever the weighting functions W^+ and W^- are not equal to each other. In particular, tailored proper scoring rules with positive scores seems to result in stronger deviations from honest reporting and, consequently, risk neutrality than with negative scores, a fact that is empirically plausible (Wakker, 2010, page 264). Furthermore, the above example illustrates that agents overweight low probabilities by reporting probability values greater than their true beliefs, and they underweight high probabilities by reporting probability values less than their true beliefs.

6 Using tailored proper scoring rules to elicit an agent's weighting functions

The results from the previous section are negative in nature because they mean that RDU agents report biased beliefs under tailored proper scoring rules. On the positive side, I discuss in this section how tailored proper scoring rules can elicit weighting functions in a parameter-free manner.

My approach assumes that there are two exhaustive and mutually exclusive outcomes, σ_1 and σ_2, with known, objective probability values ϕ and $1 - \phi$. For example, σ_1 and σ_2 can be the outcomes "heads" and "tails" in an experiment where a biased coin with known Bernoulli distribution is tossed. An agent is then asked to report his belief $\boldsymbol{\mu} = (\mu, 1 - \mu)$, for $\mu \in [0, 1]$.

Consider a proper scoring rule $R(\boldsymbol{\mu}, \sigma_x)$, for $x \in \{1, 2\}$, defined as follows:

$$R(\boldsymbol{\mu}, \sigma_1) = R'(\boldsymbol{\mu}, \sigma_1)$$
$$R(\boldsymbol{\mu}, \sigma_2) = R'(\boldsymbol{\mu}, \sigma_2) + sgn(m) \times m$$

where R' is a bounded proper scoring rule, i.e., a proper scoring rule where all the returned scores are real numbers, sgn is the sign function, and m is the maximum score returned by R'. Then, by construction, $R(\boldsymbol{\mu}, \sigma_2) \geq R(\boldsymbol{\mu}, \sigma_1)$ for any $\boldsymbol{\mu}$, which means that R is *comonotonic* (Kothiyal et al., 2011). I now construct a tailored proper scoring rule to elicit $\boldsymbol{\mu}$, i.e., $S(\boldsymbol{\mu}, \sigma_x) = U^{-1}(R(\boldsymbol{\mu}, \sigma_x))$. Since U^{-1} is strictly increasing, I then obtain $S(\boldsymbol{\mu}, \sigma_2) \geq S(\boldsymbol{\mu}, \sigma_1)$ for any $\boldsymbol{\mu}$, which implies that S also satisfies comonotonicity.

In previous sections, for ease of exposition and mathematical notation, I assumed that $S(\mathbf{q}, \theta_n) \geq S(\mathbf{q}, \theta_{n-1}) \geq \cdots \geq S(\mathbf{q}, \theta_1)$. I claimed that such an assumption is without loss of generality because the outcomes could always be rearranged *a posteriori*. In this section, however, I do not allow the outcomes to be rearranged and, by construction, $S(\boldsymbol{\mu}, \sigma_2) \geq S(\boldsymbol{\mu}, \sigma_1)$ for any belief $\boldsymbol{\mu}$. For example, in the aforementioned coin experiment, one agent will always receive higher scores if outcome σ_2 = "tails" occurs than if outcome σ_1 = "heads" occurs, no matter what the agent reports.

First, consider the case where the resulting tailored proper scoring rule S is negative, i.e., $0 \geq S(\boldsymbol{\mu}, \sigma_2) \geq S(\boldsymbol{\mu}, \sigma_1)$. Proposition 2 implies that an agent who maximizes a rank-dependent utility reports the probability vector $\boldsymbol{\mu} = (\pi_1^-, \pi_2^-) = (W^-(\phi), 1 - W^-(\phi))$. In other words, I obtain the value of W^- for the objective probability value ϕ. For a sufficiently dense set of objective probabilities, e.g., taking all values in the set $\{0, 0.05, 0.1, \ldots, 0.95, 1\}$, I obtain a parameter-free estimate of the weighting function W^-.

Alternatively, if S is a positive tailored proper scoring rule, Proposition 1 says that an agent who maximizes a rank-dependent utility reports the probability vector $\boldsymbol{\mu} = (\pi_1^+, \pi_2^+) = (1 - W^+(1 - \phi), W^+(1 - \phi))$. Then, for a sufficiently dense set of objective probabilities, I obtain a parameter-free estimate of the weighting function W^+.

Finally, if S is a mixed tailored proper scoring rule, Proposition 3 says that an agent who behaves so as to maximize a rank-dependent utility reports the probability vector $\boldsymbol{\mu} = (\pi_1^-, \pi_2^+) = (W^-(\phi), W^+(1 - \phi))$, under the assumption that $W^-(\phi) + W^+(1 - \phi) = 1$. Then, for a sufficiently dense set of objective probabilities, I obtain a parameter-free estimate of both the weighting function W^- and the weighting function W^+.

It is noteworthy that without the comonotonicity property, π_1^- is always less than or equal to 0.5, and π_2^+ is always greater than or equal to 0.5 (Kothiyal et al., 2011). Consequently, the weighting function W^- could not be estimated for probability values greater than 0.5, whereas the weighting function W^+ could not be estimated for probability values less than 0.5.

On a final note, I observe that traditional methods for eliciting weighting functions assume that agents report indifferences between lotteries honestly (Abdellaoui, 2000). Under my approach, on the other hand, it is in the best interest of an agent to report $\boldsymbol{\mu}$ honestly since this maximizes his rank-dependent utility.

7 Obtaining true beliefs from vectors of decision weights

In Section 5, I showed how tailored proper scoring rules elicit vectors of decision weights from agents who behave so as to maximize a rank-dependent utility. In Section 6, I discussed how to use tailored proper scoring rules to elicit an agent's weighting functions. A natural question that then arises regards how to combine these two results in order to obtain an agent's true belief $\mathbf{p}$ when that agent reports a vector of decision weights. In the following subsections, I show how an agent's true belief can be obtained by using simple recursive procedures. The proposed procedures are sound as long as $S(\mathbf{q}, \theta_n) > S(\mathbf{q}, \theta_{n-1}) > \cdots > S(\mathbf{q}, \theta_1)$, i.e., when there are only inequalities in the scores from the tailored proper scoring rule. Otherwise, the underlying proper scoring rule might have to satisfy comonotonicity (Kothiyal et al., 2011).

7.1 Positive tailored proper scoring rule

If a positive, tailored proper scoring rule is used in the elicitation process, then Proposition 1 says that the belief $\mathbf{q} = (q_1, \ldots, q_n) = (\pi_1^+, \ldots, \pi_n^+)$ is reported by a rank-dependent utility maximizer, which implies that:

$$W^+(p_n) = q_n$$
$$W^+(p_{n-1} + p_n) = q_{n-1} + q_n$$
$$\vdots$$
$$W^+\left(\sum_{x=2}^{n} p_x\right) = \sum_{x=2}^{n} q_x$$

Once W^+ is known, the above system of equations can be solved by using backward substitution, i.e., by first computing p_n, then substituting that into the next equation to find p_{n-1}, and so on. Starting with the base case p_n, I have $p_n = W^{+^{-1}}(q_n)$. For p_{n-1}, I have $p_{n-1} = W^{+^{-1}}(q_{n-1} + q_n) - p_n$. More generally, for all $k \in \{2, \ldots, n-1\}$, I obtain p_k by solving the equation $p_k = W^{+^{-1}}\left(\sum_{x=k}^{n} q_x\right) - \sum_{x=k+1}^{n} p_x$. Finally, $p_1 = 1 - \sum_{x=2}^{n} p_x$.

7.2 Negative tailored proper scoring rule

If a negative, tailored proper scoring rule is used in the elicitation process, then Proposition 2 says that the belief $\mathbf{q} = (q_1, \ldots, q_n) = (\pi_1^-, \ldots, \pi_n^-)$ is reported by a rank-dependent utility maximizer, which implies that:

$$W^-\left(\sum_{x=1}^{n-1} p_x\right) = \sum_{x=1}^{n-1} q_x$$
$$\vdots$$
$$W^-(p_1 + p_2) = q_1 + q_2$$
$$W^-(p_1) = q_1$$

Once W^- is known, the above system of equations can be solved by using forward substitution, i.e., by first computing p_1, then substituting that into the next equation to find p_2, and so on. Starting with the base case p_1, I have $p_1 = W^{-^{-1}}(q_1)$. For p_2, I have $p_2 = W^{-^{-1}}(q_1 + q_2) - p_1$. More generally, for all $k \in \{2, \ldots, n-1\}$, I obtain p_k by solving the equation $p_k = W^{-^{-1}}\left(\sum_{x=1}^{k} q_x\right) - \sum_{x=1}^{k-1} p_x$. Finally, $p_n = 1 - \sum_{x=1}^{n-1} p_x$.

7.3 Mixed tailored proper scoring rule

Finally, if a mixed, tailored proper scoring rule is used in the elicitation process, then Proposition 3 says that the belief $\mathbf{q} = (q_1, \ldots, q_n) = \left(\pi_1^-, \ldots, \pi_{i-1}^-, \pi_i^+, \ldots, \pi_n^+\right)$ is reported under the assumption that $W^+(\rho) + W^-(1-\rho) = 1$, for all $\rho \in [0, 1]$, which implies that:

$$W^+(p_n) = q_n$$
$$W^+(p_{n-1} + p_n) = q_{n-1} + q_n$$
$$\vdots$$
$$W^+\left(\sum_{x=i}^{n} p_x\right) = \sum_{x=i}^{n} q_x$$
$$W^-\left(\sum_{x=1}^{i-1} p_x\right) = \sum_{x=1}^{i-1} q_x$$
$$\vdots$$
$$W^-(p_1 + p_2) = q_1 + q_2$$
$$W^-(p_1) = q_1$$

Once the weighting functions W^+ and W^- are known, the above system of equations can be solved by using forward and backward substitution, i.e., forward substitution can be used to obtain the values of $p_1, \ldots, p_{i-1}$ as discussed in Section 7.2, whereas backward substitution can be used to obtain the values of $p_i, \ldots, p_n$ as discussed in Section 7.1.

8 Conclusion

Proper scoring rules are traditional devices to elicit beliefs over uncertain outcomes. As discussed in this paper, reporting a belief under a proper scoring rule is equivalent to making a decision under uncertainty. An implicit assumption when eliciting beliefs using proper scoring rules is that the underlying agents are risk neutral. Such an assumption is hardly compelling when the agents are human beings. Winkler (1969) suggested how to adapt proper scoring rules to expected utility theory by tailoring the proper scoring rule to an agent's nonlinear utility function. Currently, there is overwhelming evidence that rank-dependent models are more accurate when describing and predicting human beings' decisions under uncertainty than expected utility theory. In this paper, I characterized how an agent who maximizes a rank-dependent utility reports his belief under a tailored proper scoring rule. In particular, I found that such an agent misreports his true belief by reporting a vector of decision weights.

Decision weights can be seen as a cognitive bias concerning how human beings deal with probabilities and, thus, they should not be taken as a measure of an agent's true belief. Hence, my findings highlight the necessity of knowing all the components that drive an agent's attitude towards uncertainty before appropriately using a proper scoring rule to elicit that agent's belief.

On the positive side, I showed how to elicit weighting functions using tailored proper scoring rules, and how to obtain an agent's true belief from his misreported belief once his weighting functions are known. My work thus provides guidelines for appropriately using proper scoring rules under the empirically plausible assumption that agents behave so as to maximize rank-dependent utilities. The first step consists of eliciting the agent's utility function, e.g., by using the tradeoff method proposed by Wakker and Deneffe (1996). In the second step, the agent's utility function is incorporated into a proper scoring rule, and the resulting tailored proper scoring rule is used to elicit the agent's belief. In the third step, the agent's weighting functions are elicited using tailored proper scoring rules, as described in Section 6. Finally, the agent's true belief is obtained *a posteriori* from his misreported belief, as described in Section 7. This approach is rather general in a sense that it works for any strictly increasing utility function, any finite number of outcomes, and any proper scoring rule as long as the potential scores given a reported belief are all different from each other.

It is interesting to note that the analysis performed in this paper can be extended to other non-expected utility theories. For example, consider the rank-affected multiplicative weights (RAM) model by Birnbaum (1997, 2008). For two outcomes, θ_1 and θ_2, and a positive, tailored proper scoring rule $S(\mathbf{q}, \theta_x)$, where the outcomes are ordered such that $S(\mathbf{q}, \theta_2) \geq S(\mathbf{q}, \theta_1)$, the RAM model is:

$$\frac{2 \times p_1^\gamma \times U(S(\mathbf{q}, \theta_1))}{2 \times p_1^\gamma + 1 \times p_2^\gamma} + \frac{1 \times p_2^\gamma \times U(S(\mathbf{q}, \theta_2))}{2 \times p_1^\gamma + 1 \times p_2^\gamma} \quad (9)$$

Intuitively, the RAM model means that the value an agent assigns to a lottery is equal to a weighted average in which the weight associated with a payoff is a function of the probability associated with the underlying outcome and the rank of the payoff relative to other payoffs. Instead of his true belief $\mathbf{p} = (p_1, p_2)$, an agent who behaves so as to maximize the above function ends up reporting the following belief:

$$\mathbf{q} = \left(\frac{2 \times p_1^\gamma}{2 \times p_1^\gamma + 1 \times p_2^\gamma}, \frac{1 \times p_2^\gamma}{2 \times p_1^\gamma + 1 \times p_2^\gamma} \right)$$

For example, consider the true belief $\mathbf{p} = (0.2, 0.8)$ used in the numerical example in Section 5.1. Moreover, assume the parameter value $\gamma = 0.7$ in (9). In this setting, in order to maximize (9), an agent reports $\mathbf{q} = (0.431, 0.569)$. Note that the reported belief is different than $(0.393, 0.607)$ and $(0.257, 0.743)$, the beliefs reported under RDU for, respectively, a positive and a negative tailored proper scoring rules (see Section 5.1).

As can be seen from the above example, different decision theories might imply different reporting behavior under proper scoring rules. Consequently, the procedure to obtain an agent's true belief from his reported belief is also dependent on the underlying decision theory. These points raise an important question: *which decision theory is the "correct" theory when eliciting beliefs using proper scoring rules*? Identifying the "best theory" naturally requires judgments about the relative importance of predictive accuracy, simplicity, tractability, theoretical properties, etc. Such judgments are often subjective in their nature. For example, one might argue that rank-dependent models have stronger axiomatic foundations in terms of preferences than the RAM model. Alternatively, the RAM model accounts for behavior that many rank-dependent models violate, such as coalescing and violations of stochastic dominance (Birnbaum, 2008).

Another example of such a trade-off concerns the Transfer of Attention Exchange (TAX) model by Birnbaum and Chavez (1997). Birnbaum (2008) documented the superior predictive performance of the TAX model over some rank-dependent models as well as the RAM model. The TAX model represents the utility of a lottery as a weighted average of the utilities of payoffs, where the weights depend on both the probabilities of the outcomes and the ranks of the payoffs. Unlike weights in rank-dependent models, those weights represent transfers of attention from branch to branch. In practice, this implies that the utility of each payoff is weighted by a nonlinear transformation of a subjective probability as well as "weight transfer" factors. Such factors make the problem of adapting proper scoring rules to the general TAX model quite challenging, a task that I have not been able to accomplish yet.

If one decides that predictive accuracy is the most relevant criterion, then a whole new set of experiments might be required to determine the most appropriate decision theory when using proper scoring rules. As discussed in Section 4.2, the payoffs of lotteries are defined by an agent's reported belief when using proper scoring rules. Consequently, agents have some control over their payoffs. In practice, this fact might have some influence on the way agents choose amongst different lotteries.

References

Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science*, *46*(11), 1497–1512.

Allais, M. (1953). Violations of the betweenness axiom and nonlinearity in probability. *Econometrica*, *21*, 503-546.

Arrow, K. J. (1971). *Essays in the theory of risk-baring* (Vol. 1). Markham Publishing Company Chicago.

Bacon, D. F., Chen, Y., Kash, I., Parkes, D. C., Rao, M., & Sridharan, M. (2012). Predicting your own effort. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems* (pp. 695–702).

Birnbaum, M. H. (1997). Violations of monotonicity in judgment and decision making. In A. A. J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. duncan luce* (p. 73-100).

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological review*, *115*(2), 463–501.

Birnbaum, M. H., & Chavez, A. (1997). Tests of theories of decision making: Violations of branch independence and distribution independence. *Organizational Behavior and Human Decision Processes*, *71*(2), 161–194.

Camerer, C. F. (2004). Prospect theory in the wild : Evidence from the field. In C. F. Camerer, G. Loewenstein, & M. Rabin (Eds.), *Advances in behavioral economics* (pp. 148–161).

Carvalho, A., & Larson, K. (2010). Sharing a reward based on peer evaluations. In *Proceedings of the 9th international conference on autonomous agents and multiagent systems* (pp. 1455–1456).

Carvalho, A., & Larson, K. (2011). A truth serum for sharing rewards. In *Proceedings of the 10th international conference on autonomous agents and multiagent systems* (pp. 635–642).

Carvalho, A., & Larson, K. (2012). Sharing rewards among strangers based on peer evaluations. *Decision Analysis*, *9*(3), 253–273.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive psychology*, *38*(1), 129–166.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Johnstone, D. J. (2011). Economic interpretation of probabilities estimated by maximum likelihood or score. *Management Science*, *57*(2), 308–314.

Johnstone, D. J., Jose, V. R. R., & Winkler, R. L. (2011). Tailored scoring rules for probabilities. *Decision Analysis*, *8*, 256–268.

Jose, V. R. (2009). A characterization for the spherical scoring rule. *Theory and Decision*, *66*(3), 263–281.

Kothiyal, A., Spinu, V., & Wakker, P. P. (2011). Comonotonic proper scoring rules to measure ambiguity and subjective beliefs. *Journal of Multi-Criteria Decision Analysis*, *17*(3-4), 101–113.

Nakazono, Y. (2013). Strategic behavior of federal open market committee board members: Evidence from members' forecasts. *Journal of Economic Behavior & Organization*, *93*, 62–70.

Offerman, T., Sonnemans, J., Van De Kuilen, G., & 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.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*(4), 323–343.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783–801.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, *57*(3), 571–587.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, *1*(1), 43–62.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, 332–382.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? how can we know?* Princeton University Press.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge University Press.

Wakker, P. P., & Deneffe, D. (1996). Eliciting von neumann-morgenstern utilities when probabilities are distorted or unknown. *Management Science*, *42*(8), 1131–1150.

Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, *64*(327), 1073–1078.

Winkler, R. L., & Murphy, A. H. (1968). "good" probability assessors. *Journal of Applied Meteorology*, *7*(5), 751–758.

Winkler, R. L., & Murphy, A. H. (1970). Nonlinear utility and the probability score. *Journal of Applied Meteorology*, *9*, 143–148.