A TRUTH-SERUM FOR NON-BAYESIANS: CORRECTING PROPER SCORING RULES FOR RISK ATTITUDES[1]

T. Offerman[a], J. Sonnemans[a], G. van de Kuilen[b], and P.P. Wakker[c]

a: *CREED, Dept. of Economics, University of Amsterdam, Roetersstraat 11, Amsterdam, 1018 WB, The Netherlands*

b: *TIBER, Dept. of Economics, Tilburg University, P.O. Box 90153, Tilburg, 5000 LE, The Netherlands*

c: *Econometric Institute, Erasmus University, P.O. Box 1738, Rotterdam, 3000 DR, the Netherlands*

December, 2008

Review of Economic Studies, forthcoming

ABSTRACT. Proper scoring rules provide convenient and highly efficient tools for incentive-compatible elicitations of subjective beliefs. As traditionally used, however, they are valid only under expected value maximization. This paper shows how they can be generalized to modern ("nonexpected utility") theories of risk and ambiguity, yielding mutual benefits: users of scoring rules can benefit from the empirical realism of nonexpected utility, and analysts of ambiguity attitudes can benefit from efficient measurements using proper scoring rules. An experiment demonstrates the feasibility of our generalization.

KEY WORDS: *belief measurement, proper scoring rules, ambiguity, Knightian uncertainty, subjective probability, nonexpected utility*

JEL-CLASSIFICATION: D81, C60, C91

[1] A preliminary version of this paper circulated with the title "Is the Quadratic Scoring Rule Really Incentive Compatible?"

1. INTRODUCTION

An important problem in mechanism design concerns the elicitation of private information. A design is incentive compatible if the actions of agents, motivated solely by self-interest, nevertheless reveal their true private information (Hurwicz 1960). A social planner can then use all relevant information to devise the most efficient social allocation. This paper considers the case where the private information concerns subjective beliefs about the likelihood of uncertain events, often modelled using subjective probabilities. This case arises, for instance, when principals rely on the judgments of specialized agents. In the absence of proper incentives, agents may pretend to be more confident about their judgment than they really are, and may not update their beliefs sufficiently, so as to suggest greater ability than they really have (Li 2007). Manski (2004) presented an historical survey of belief measurement, and gave many economic applications.

For belief measurement, incentive-compatible mechanisms had been known at an early stage in the form of proper scoring rules (Brier 1950; Good 1952). These scoring rules are particularly efficient mechanisms for eliciting subjective beliefs in an incentive-compatible manner. They use cleverly designed optimization problems where the observation of one single choice suffices to determine the exact quantitative degree of belief of an agent in an uncertain event. Hence, they have recently become popular in experimental economics and game theory (Nyarko and Schotter 2002). Proper scoring rules have been used in many other fields in the social sciences, including accounting (Wright 1988), Bayesian statistics (Savage 1971), business (Staël von Holstein 1972), education (Echternacht 1972), finance (Johnstone 2007a,b; Shiller, Kon-Ya, and Tsutsui 1996), medicine (Spiegelhalter 1986), meteorology (Palmer and Hagedorn 2006; Yates 1990), politics (Tetlock 2005), psychology (McClelland and Bolger 1994), and other fields (Hanson 2002; Prelec 2004).

An alternative way to elicit beliefs that has recently become popular concerns prediction markets on the internet (Wolfers and Zitzewitz 2004). Here people trade event-contingent payments regarding uncertain events, such as a guarantee to receive €100 if a Democrat becomes the next President of the United States. If this guarantee is now traded at a price P, then P/100 is taken as the market-probability of the event. This inference assumes expected value. Johnstone (2007a) explained that a financial market can, for many purposes, be analyzed as if it were a rational individual, and discussed the role of proper scoring rules in

such settings. In a market with agents who maximize the logarithmic utility of wealth, good forecasters are better identified by their performance in terms of proper scoring rules than in terms of their average earnings (Johnstone 2007b).

Whereas all applications of proper scoring rules that we are aware of assume expected value maximization ("risk neutrality"), many deviations have been observed empirically. Under expected utility, risk aversion is the common finding (Bernoulli 1738). Johnstone (2007a) and Winkler and Murphy (1970) discussed the implications of risk aversion for proper scoring rules. Further, many deviations from expected utility have been found, both when probabilities exist ("risk": Allais 1953; Kahneman and Tversky 1979) and when probabilities cannot even be specified ("ambiguity": Ellsberg 1961; Keynes 1921; Knight 1921).

This paper extends proper scoring rules from the expected value model as assumed in the 1950s, when proper scoring rules were introduced, to the current state of the art in decision theory. Thus we can, on the one hand, improve the validity of belief measurement using proper scoring rules. On the other hand, we can use proper scoring rules to obtain more efficient methods for measuring risk and ambiguity attitudes. In economics, probabilities are usually not known, and the importance of quantitative measurements of ambiguity attitudes has been widely understood (Gilboa 2004; Greenspan 2004). We show how subjective beliefs and ambiguity attitudes can be isolated from risk attitude in a surprisingly easy way by means of proper scoring rules. Thus, we can correct measurements of subjective beliefs and ambiguity attitudes for nonneutral risk attitudes.

We illustrate the feasibility of our method by an experiment where we measure the subjective beliefs of participants about the future performance of stocks after the provision of information about past performance. The empirical findings confirm the usefulness of our method. However, violations of additivity of subjective beliefs are reduced but not eliminated by our corrections. Thus, the classical measurements will contain violations of additivity that are partly due to the incorrect assumption of expected value maximization, but partly they are genuine. Subjective beliefs are genuinely nonadditive.

The analysis in this paper consists of three parts. The first part (§§2-4) considers various modern theories of risk and ambiguity, and derives implications for proper scoring rules. The second part (§§5-6) applies the revealed-preference technique to the results of the first part. That is, we do not assume theoretical models to derive implications for empirical observations, but we use empirical observations to derive implications for theoretical models. §5 presents the main result of this paper, showing how subjective beliefs can easily be derived from observed choices using what are called risk-corrections. §6 gives an example to

illustrate such a derivation at the individual level. Readers who are only interested in applying our method empirically can skip most of §§3-5, only reading Corollary 5.4. A short direct proof of this result, showing that it holds for general proper scoring rules, is given following the corollary.

The third part of the paper (§§7-11) implements our correction method in an experiment. In order to demonstrate the applicability of our method, we use it to investigate some properties of nonadditive beliefs and of different implementations of real incentives. §7 contains methodological details. §8 presents the results regarding the biases that we correct for, and §9 explains some implications of the corrections of such biases. §10 provides an additional control treatment. The experimental results are discussed in §11. A general discussion and conclusions are in §§12-13. Appendix A makes some technical remarks, Appendix B presents the proofs, and Appendix C surveys the implications of modern decision theories for our measurements. The experimental instructions are in Appendix D.

2. PROPER SCORING RULES; DEFINITIONS

Let E denote an event such that an agent is uncertain about whether or not the event obtains, such as whether a stock's value will decrease during the next six months. The degree of uncertainty of the agent about E will obviously depend on the information that the agent possesses about E. For most uncertain events, no objective probabilities of occurrence are known, and decisions have to be based on subjective likelihood assessments.

Prospects refer to event-contingent payments. We use the general notation $x_E y$ for a *prospect* that yields outcome x if event E obtains and outcome y if E^c obtains, with E^c the *complementary event* not-E. *Outcomes* are money amounts. *Risk* concerns the special case of known probabilities. Then, for a prospect $x_E y$, the probability p of event E is known. We identify this prospect with a probability distribution $x_p y$ over money, yielding x with probability p and y with probability 1−p.

This paper considers the *quadratic scoring rule* (*QSR*), the most commonly used proper scoring rule (McKelvey and Page 1990; Nyarko and Schotter 2002; Palfrey and Wang 2007). A *QSR-prospect*

$$(1-(1-r)^2)_E(1-r^2) \tag{2.1}$$

is offered to the agent, where $0 \le r \le 1$ is chosen at the agent's discretion. This number r is a function of E, sometimes denoted r_E, and is called the (*uncorrected*) *reported probability* of E. The reasons for using this term will be explained later. If event E has an (objective or subjective) probability p, then, according to all theories considered, r_E will depend only on p, so that we can write it as a function R(p).

Instead of Eq. 2.1, we could have used more general prospects $(a–b(1–r)^2)_E(a–br^2)$ for any $b > 0$ and $a \in \mathbb{R}$. For simplicity, we restrict our attention to $a = b = 1$ as in Eq. 2.1. No negative payments can occur then, so that the agent never loses money. Under the event that happens, the QSR in fact yields 1 minus the squared distance between the reported probability of a clairvoyant (who assigns probability 1 to the event that happens) and the reported probability of the agent (this probability is r under E, and 1–r under E^c). The following observation about a symmetry of the QSR will be useful.

OBSERVATION 2.1. The quadratic scoring rule for event E presents the same choice of prospects as the quadratic scoring rule for event E^c, with each prospect resulting from r as the reported probability of E identical to the prospect resulting from 1–r as the reported probability of E^c. □

Because of Observation 2.1, we have

$$r_{E^c} = 1– r_E \tag{2.2}$$

and

$$R(1–p) = 1 – R(p). \tag{2.3}$$

Hence, we will state many results only for $r \ge 0.5$. The case $r < 0.5$ then follows from Eqs. 2.2 and 2.3 applied to E^c.

3. A THEORETICAL ANALYSIS OF PROPER SCORING RULES

In this section we consider modern decision models for decision making under uncertainty, and derive implications for proper scoring rules. As explained in detail in Appendix C, virtually all currently existing models, including multiple priors (Gilboa and

Schmeidler 1989) and Choquet expected utility (Gilboa 1987; Schmeidler 1989) evaluate the QSR-prospect of Eq. 2.1 using the following formula.

$$\text{For } r \geq 0.5\text{: } W(E)U(1-(1-r)^2) + (1-W(E))U(1-r^2). \qquad (3.1)$$

Comments on the case $r < 0.5$ follow later. U is the *utility function*, assumed to be continuous and strictly increasing, and scaled such that $U(0) = 0$. We present a number of cases for W, with each case generalizing the preceding one. Cases 1 and 2 are well known.

CASE 1 [*Expected Value*]. U is the identity function and W is a probability measure P.

CASE 2 [*Expected Utility*]. W is a probability measure P.

CASE 3 [*Probabilistic Sophistication* (with nonexpected utility)]. There exist a probability measure P and a continuous strictly increasing function w, the *probability weighting function*, such that $W(.) = w(P(.))$, $w(0) = 0$, and $w(1) = 1$.

CASE 4 [*General Model*]. W satisfies: (i) $W(\varnothing) = 0$; (ii) $W = 1$ for the universal event; (iii) $C \supset D$ implies $W(C) \geq W(D)$.

We distinguish two subcases for Case 3 and, hence, also for Cases 2 and 1.

SUBCASE a. [*Objective Probabilities*]. The probability measure P is objective, based on statistical data that everyone agrees on.

SUBCASE b. [*Subjective Probabilities*]. The probability measure P may be subjective, and can be revealed from preferences.[2]

De Finetti (1937), Savage (1954), and Machina and Schmeidler (1992) gave preference foundations for Cases 1.b, 2.b, and 3.b. Case 3 is an interesting intermediate case, with the Bayesian principles violated at the level of decisions but not at the level of beliefs. In the general Case 4, the Bayesian principles are also violated at the level of beliefs. The well-known Allais (1953) paradox shows that expected utility is often violated, so that w and W

[2] In this paper, the term *subjective probability* is used only for probability judgments that are Bayesian in the sense that they satisfy the laws of probability. In the literature, the term subjective probability has sometimes been used for judgments that deviate from the laws of probability, including cases where these judgments are nonlinear transformations of objective probabilities when the latter are given. We use the term (probability) weights or beliefs, depending on the way of generalization, to refer to the latter.

are nonadditive and we cannot restrict attention to classical Cases 1 and 2. The well-known Ellsberg (1961) paradox, discussed in detail later, shows that probabilistic sophistication is often violated, so that the general Case 4 has to be considered.

For the general model, the formula to evaluate the QSR-prospect of Eq. 2.1 for $r < 0.5$ follows from Observation 2.1:

$$\text{For } r < 0.5\text{: } (1-W(E^c))U(1-(1-r)^2) + W(E^c)U(1-r^2). \tag{3.2}$$

For expected value and expected utility, Eq. 3.2 agrees with Eq. 3.1 and the two formulas can be used interchangeably, but for probabilistic sophistication and the general model, Eq. 3.2 can be different. The latter separate, "rank-dependent," way of weighting the outcomes, with weights always summing to 1, was discovered independently by Quiggin (1982) for the special case of risk with given probabilities, and by Schmeidler (1989; first version 1982) for the general model. This idea was the key to the development of the modern nonexpected utility theories. It was incorporated in the new version of prospect theory (Tversky and Kahneman 1992).

Objective probabilities can best be interpreted as a special limiting case of subjective probabilities, a point formalized by Machina (2004). The hypothetical situation of an agent using a subjective probability different from an objective probability, if the latter is given, cannot arise under plausible assumptions (Wakker 2009).

We now analyze which optimal values r_E are predicted under the various cases considered.

THEOREM 3.1. In the general model, the optimal choice r in Eq. 2.1 satisfies:

$$\text{If } r > 0.5, \text{ then } r = r_E = \frac{W(E)}{W(E) + (1-W(E))\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}. \tag{3.3}$$

□

The optimality result for $r < 0.5$ follows from Observation 2.1 applied to E^c. If $r = 0.5$ is optimal then it can be a boundary solution for which Eq. 3.3 need not hold. We will discuss this case later. Theorem 3.1 generalizes results, obtained by Winkler and Murphy (1970) for expected utility, to general nonexpected utility. The following corollary, first found by Brier

(1950), is highly appealing, and is, to the best of our knowledge, the first incentive-compatible result provided in the literature.

COROLLARY 3.2. Under expected value, Eq. 3.3 holds for all r and $r = r_E = P(E)$. □

Thus, under expected value, it is in the agent's best interest to report true subjective probabilities. The following section illustrates the extent to which reported probabilities, still commonly equated with subjective probability in virtually all applications today, can deviate from subjective probabilities because of empirical deviations from expected value. We next consider the case r = 0.5 under expected utility.

OBSERVATION 3.3. Under expected utility with probability measure P, $r_E = 0.5$ implies $P(E) = 0.5$. Conversely, $P(E) = 0.5$ implies $r_E = 0.5$ if risk aversion holds. Under risk seeking, $r_E \neq 0.5$ is possible if $P(E) = 0.5$. □

4. DISCREPANCIES BETWEEN SUBJECTIVE PROBABILITIES AND PROPER SCORING RULES: NUMERICAL EXAMPLES

The solutions r presented in this section can be verified by substitution in the implicit Eq. 3.3. We will later provide explicit expressions for $R^{-1}(p)$, which we used to find the solutions and to draw Figure 4.1. We consider two urns each containing 100 balls that are Crimson, Green, Silver, or Yellow. Urn K ("known") contains 25 balls of each colour, and urn A ("ambiguous") contains the balls in an unknown proportion. One ball will be drawn at random from each urn. C denotes the event of a crimson ball drawn from urn K, with G, S, and Y defined similarly. E is the event that the ball drawn from K is not crimson, i.e. it is the event $C^c = \{G,S,Y\}$. C_a denotes the event of a crimson ball drawn from urn A, with G_a, S_a, and Y_a defined similarly, and $E_a = C_a^c$. Subjects are asked to report their belief in event E and are rewarded with a QSR (Eq. 2.1). We now consider the four cases presented in the preceding section.

CASE 1 [*Expected Value*]. Expected value holds for urn K. Then $r_{E_K} = R(0.75) = 0.75$ is optimal in Eq. 2.1. The point r_E is depicted as r^{EV} in Figure 4.1, at p = 0.75. Corollary 3.2 implies that $r_G = r_S = r_Y = 0.25$. The reported probabilities satisfy additivity: $r_G + r_S + r_Y = r_E$.

CASE 2 [*Expected Utility*]. Expected utility holds for urn K, with $U(x) = x^{0.5}$. We obtain $r_E = R(0.75) = 0.69$, depicted as r^{EU} in Figure 4.1, at p = 0.75. The expected value of the resulting QSR-prospect is 0.0031 (i.e. 0.8125 – 0.8094) less than it was in Case 1. This difference can be interpreted as a risk premium, designating a profit margin for an insurance company. By Eq. 2.2, $r_C = 0.31$, and by symmetry $r_G = r_S = r_Y = 0.31$ too. The reported probabilities violate additivity, with $r_G + r_S + r_Y = 0.93 > 0.69 = r_E$. Because of this violation, the data can directly reveal that expected value, the common assumption in applications of proper scoring rules, does not hold. □

CASE 3 [*Nonexpected Utility with Probabilistic Sophistication*]. Probabilistic sophistication holds for urn K, with $U(x) = x^{0.5}$, and

$$w(p) = \left(exp(-(-ln(p))^{\alpha})\right) \tag{4.1}$$

with parameter $\alpha = 0.65$ (Prelec 1998). This function agrees with the prevailing empirical findings (Tversky and Kahneman 1992; Abdellaoui 2000; Bleichrodt and Pinto 2000; Gonzalez and Wu 1999). We obtain $r_E = R(0.75) = 0.61$, depicted as r^{nonEU} in Figure 4.1 at p = 0.75. The extra expected-value loss (and, hence, the extra risk premium) relative to Case 2 is 0.0174 (i.e. 0.8094 – 0.7920). By Eq. 2.2, $r_C = 0.39$, and, by symmetry, $r_G = r_S = r_Y = 0.39$ too. The reported probabilities strongly violate additivity because $r_G + r_S + r_Y = 1.17 > 0.61 = r_E$. □

The following case describes the most fundamental deviation from expected value and expected utility, driven by ambiguity, a central topic in decision theory today. The case concerns a version of Ellsberg's (1961) paradox.

CASE 4 [*General Case; Violation of Probabilistic Sophistication*]. We assume probabilistic sophistication for urn K but consider, in addition, urn A using the general model. If probabilities were assigned to drawings from urn A and probabilistic sophistication were also to hold for this urn, then, in view of symmetry, we should have $P(C_a) = P(G_a) = P(S_a) =$

$P(Y_a)$. Then these probabilities would be 0.25. $P(E_a)$ would then be 0.75, as was P(E) in Case 3. Under probabilistic sophistication combined with nonexpected utility as in Case 3, r_{E_a} would be the same as r_E in Case 3 for the known urn, i.e. $r_{E_a} = 0.61$. It implies that people would be indifferent between $x_E y$ and $x_{E_a} y$ for all x and y. The latter condition is, however, typically violated empirically. People usually have a strict preference for known probabilities, implying for instance

$$100_E 0 \succ 100_{E_a} 0 \ .^3 \qquad (4.2)$$

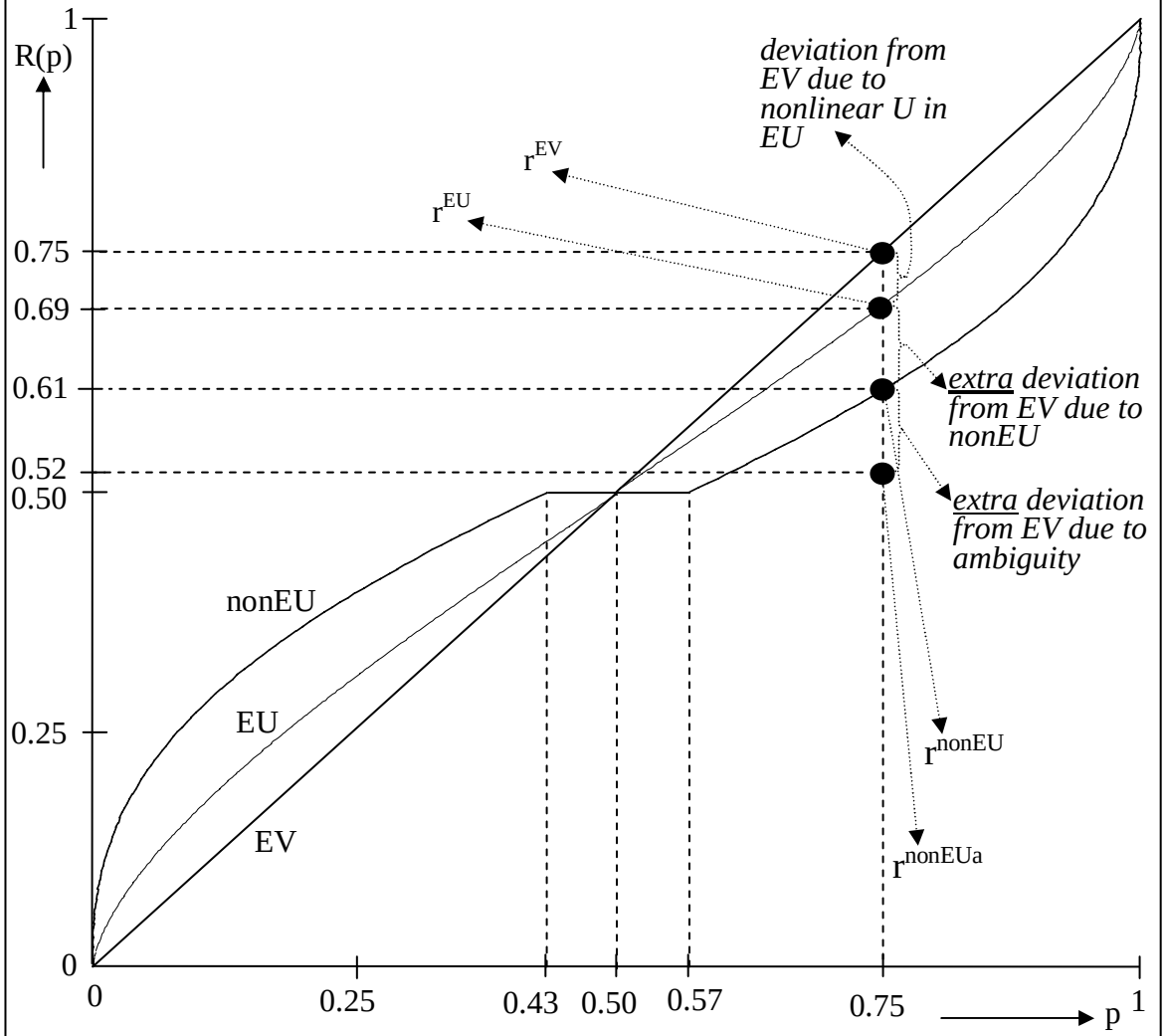
Consequently, it is impossible to model beliefs about uncertain events E_a with probabilities, and probabilistic sophistication fails. Eq. 4.2 implies that $W(E_a) < W(E)$. By Eq. 3.3, $r_{E_a} < r_E$.[4] Given the strong aversion to unknown probabilities that is often found empirically (Camerer and Weber 1992), we will assume that $r_{E_a} = 0.52$. It is depicted as r^{nonEUa} in Figure 4.1. The extra expected-value loss relative to Case 3 is 0.7920 – 0.7596 = 0.0324. This amount can be interpreted as the ambiguity-premium. By Eq. 2.2, $r_C = 0.48$, and by symmetry $r_G = r_S = r_Y = 0.48$ too. The reported probabilities violate additivity to an extreme degree, with $r_G + r_S + r_Y = 1.44 > 0.52 = r_{E_a}$. □

Figure 4.1 illustrates the extent to which reported probabilities can deviate from subjective probabilities, because of violations of expected value. The cases presented in this section concerned p = 0.75, but Figure 4.1 deals with all probabilities p under probabilistic sophistication. In the general model there are only events and no probabilities, so that the latter cannot be put on the x-axis and no graph can be drawn. For expected utility, a similar figure is in Winkler and Murphy (1970, Figure 3). Its pattern was confirmed empirically by Huck and Weizsäcker (2002). The figure illustrates the errors generated by the assumption of expected value maximization (an assumption still generally used in applications of proper scoring rules today) from the perspective of modern views in decision theory. Johnstone (2007a, p. 164) gave similar results from the perspective of a mean-variance model.

[3] This also holds if people can choose the three colours to bet on in the ambiguous urn, so that there is no reason to suspect unfavourable compositions.

[4] It is easiest to see in Eq. 3.3 that $1/r_E$ is decreasing in W(E).

FIGURE 4.1. Reported probability R(p) as a function of probability p



EV: expected value;
EU: expected utility with $U(x) = x^{0.5}$;
nonEU: nonexpected utility for known probabilities, with $U(x) = x^{0.5}$ and with w(p) as in Case 3.
nonEUa: nonexpected utility for unknown probabilities ("Ambiguity").

5. REVEALED PREFERENCE TECHNIQUES TO ELICIT SUBJECTIVE BELIEFS FROM PROPER SCORING RULES

In the preceding sections we assumed theoretical decision models and derived predictions about reported probabilities in proper scoring rules. This section presents the usual revealed-preference technique. That is, we assume that we observe reported

probabilities, and we then investigate what we can infer about decision models and their parameters. In particular, we will be interested in inferring subjective probabilities and their generalizations from proper scoring rules.

If we could observe enough general decisions under risk without proper scoring rules, with enough events available with known probabilities (such as those referring to urn K in Case 4), then we could, in principle, reveal the whole function w. Similarly, we could reveal the whole function W if we could observe enough decisions under uncertainty. Then we could obtain the following concept, which will be central in this paper.

$$B(E) = w^{-1}W(E)\ . \tag{5.1}$$

In general, B assigns value 0 to the vacuous event $\varnothing$ and value 1 to the universal event, and B is increasing in the sense that $C \supset D$ implies $B(C) \geq B(D)$. These properties similarly hold for the composition $W(\cdot) = w(B(\cdot))$, as we saw before. Under probabilistic sophistication (including expected utility and expected value), B(E) agrees with the probability P(E). In all cases in §4 up to Case 3, it is indeed the case that $B(E) = 0.75 = P(E)$. Thus, B(E) is a better candidate for measuring subjective beliefs than r_E, the value still commonly used in applications of proper scoring rules today. Whenever subjective probabilities exist, B measures them correctly, irrespective of the risk attitude. B is what results from r_E after correction for nonneutral risk attitudes. We call B the (*risk-*)*corrected reported probability*.

Case 4 showed that decisions sometimes cannot be modelled using subjective probabilities. In particular, B in Eq. 5.1 will not be a probability measure in Case 4. Yet we think that B is a better candidate to reflect subjective beliefs of likelihood than the uncorrected reported probabilities. Risk attitude is a behavioural component rather than a component reflecting beliefs and, hence, it should be filtered out from belief assessments. Many studies of direct judgments of belief have supported the thesis that subjective beliefs cannot be modelled through probabilities (McClelland and Bolger 1994; Shafer 1976; Tversky and Koehler 1994), so that B will violate additivity. Bounded rationality is an extra reason to expect that subjective beliefs will violate the laws of probability (Aragones et al. 2005; Charness and Levin 2005).

EXAMPLE 5.1. Consider Case 4. The belief component $B(E_a)$ is estimated to be $w^{-1}(W(E_a)) = w^{-1}(0.52) = 0.62$. This value implies that B must violate additivity. Under additivity, we would have $B(C_a) = 1 - B(E_a) = 0.38$, and then, by symmetry, $B(G_a) = B(S_a) = B(Y_a) = 0.38$,

so that $B(G_a) + B(S_a) + B(Y_a) = 3 \times 0.38 = 1.14$. This value should then equal $B\{G_a,S_a,Y_a\} = B(E_a) = 0.62$, but it does not. Additivity is violated and B is no probability measure.

Of the total deviation of $r_{E_a} = 0.52$ from 0.75, which is 0.23, a part of 0.06 + 0.08 = 0.14 is the result of deviations from risk neutrality that distorted the measurement of $B(E_a)$. The remaining 0.09 is not a distortion in the measurement of belief. It rather shows that belief is genuinely nonadditive. □

The measurement of B through entire measurements of w and W is laborious, in particular because of interactions with utility (Tversky and Kahneman 1992, p. 311; Abdellaoui, Vossman, and Weber 2005). The following results prepare for a tractable measurement of B. Whereas the expression of r in terms of W in Theorem 3.1 was implicit, we now present an explicit expression of its inverse, i.e. of W in terms of r. For easy later reference, we state the result for $B = w^{-1}(W)$ instead of W.

COROLLARY 5.2. For the optimal choice $r = r_E$:

$$\text{If } r > 0.5, \text{ then } B(E) = w^{-1}\left(\frac{r}{r + (1-r)\frac{U'(1-(1-r)^2)}{U'(1-r^2)}}\right). \tag{5.2}$$

□

We next display the special case of W = w(P) (with P objective or subjective), in which case $B = P = R^{-1}(r)$.

COROLLARY 5.3. Under probabilistic sophistication, we have for the optimal choice r = R(p):

$$\text{If } r > 0.5, \text{ then } p = w^{-1}\left(\frac{r}{r + (1-r)\frac{U'(1-(1-r)^2)}{U'(1-r^2)}}\right). \tag{5.3}$$

□

As it so happens, the right-hand sides in Eqs. 5.2 and 5.3 are identical. This allows a particularly convenient way to measure B.

COROLLARY 5.4. Assume the general model (Eqs. 3.1 and 3.2). For an event E with $r_E = r$ we can find the objective probability p with the same value $R(p) = r$, and then we can conclude that $B(E) = p$. That is,

$$\text{If } r_E = r = > 0.5, \text{ then } B(E) = R^{-1}(r). \tag{5.4}$$

□

The corollary is useful for empirical applications because all the terms involved are easily observable. The corollary is the only implication of our theoretical analysis that is needed for our application. Two points underlie the corollary. First, it can be applied where proper scoring rules uniquely identify the underlying subjective factors so that the inverse function R^{-1} can be defined. Second, R(p) and r_E optimize the same goal function for the case $p = B(E)$ (Eqs. 3.1 and 3.2 with $W(E) = w(B(E)) = w(p)$). Then, by the uniqueness mentioned, they must be identical. This reasoning shows that Corollary 5.4 holds for all proper scoring rules and not just for the quadratic one. It confirms that B rather than W is a non-Bayesian analog of subjective probability.

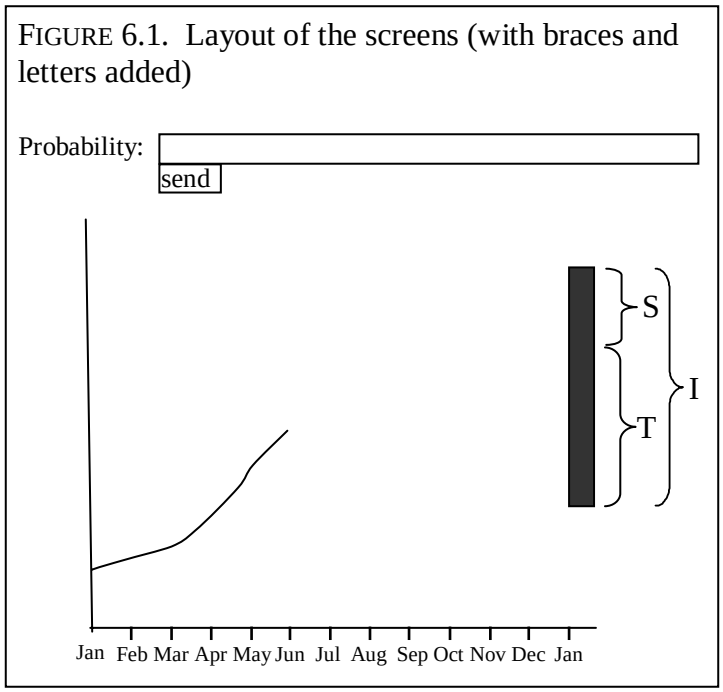
In practice, we first infer the (for the participant) optimal R(p) for a set of objective probabilities p which is so dense that we obtain a sufficiently accurate estimation of R and R^{-1}. In our experiment, we will consider all values $p = j/20$ for $j \geq 10$. Then, for all uncertain events E (or E^c if $r < 0.5$) we derive B(E) from the observed r_E using Eq. 5.4. For $r_E = 0.5$, B(E) and the inverse p may not be uniquely determined because of the flat part of R_{nonEU} in Figure 4.1. The case $r < 0.5$ follows from Eqs. 5.4 and 2.2, as always. We call the function R^{-1} the *risk-correction* (for proper scoring rules). $R^{-1}(r_E) = B(E)$ is the corrected reported probability.

6. AN ILLUSTRATION OF OUR MEASUREMENT OF BELIEF

This section describes risk corrections for a participant in the experiment in order to illustrate how our method can be applied empirically. It will show that Corollary 5.4 is the

only result of the theoretical analysis needed to apply our method. Results and curves for r < 0.5 are derived from r > 0.5 using Eq. 2.2; we will not mention this point explicitly in what follows.

FIGURE 6.1. Layout of the screens (with braces and letters added)



The left side of Figure 6.1 displays the performance of stock 12 (the Royal Begemann Group) in our experiment from January 1, until June 1, 1991, as given to the participants. Further details (such as the absence of a unit on the y-axis) will be explained in §7. The right side of the figure displays two disjoint intervals S and T, and their union I = S∪T. For each of the intervals S,T, and I, participants reported the probability of the stock ending up in that interval on January 1, 1992 (with some other questions in between the three questions considered here). For participant 14, the results are as follows.

$$r_S = 0.35;\ r_T = 0.55;\ r_I = 0.65. \tag{6.1}$$

Under additivity of reported probability, $r_S + r_T - r_I$ (the *additivity bias*, defined in general following Eq. 7.5), should be 0, but here it is not and additivity is violated.

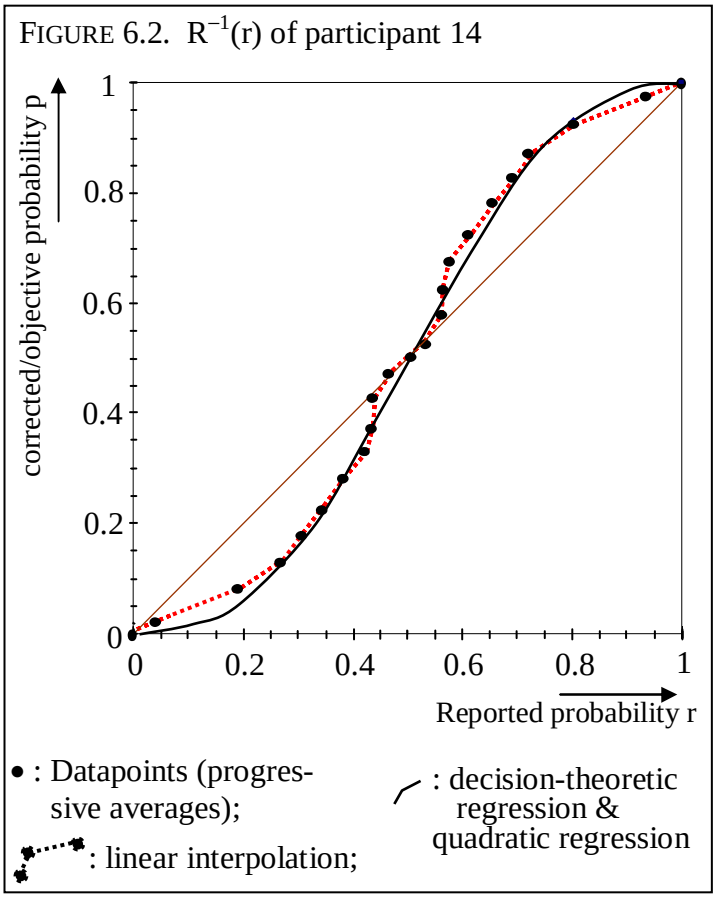
$$\text{The additivity bias is } 0.35 + 0.55 - 0.65 = 0.25. \tag{6.2}$$

Table 6.1 and Figure 6.2 (in inverted form) display the reported probabilities that we measured from this participant as a function of objective probabilities, with the curves

explained later. We use progressive averages (midpoints between data points) so as to reduce noise.[5]

TABLE 6.1. Progressive average reported probabilities R(p) of participant 14

P	.025	.075	.125	.175	.225	.275	.325	.375	.425	.475	.525	.575	.625	.675	.725	.775	.825	.875	.925	.975
R(p)	.067	.192	.267	.305	.345	.382	.422	.435	.437	.470	.530	.563	.565	.578	.618	.655	.695	.733	.808	.933



For simplicity of presentation, we analyze the data here using linear interpolation. Then R(0.23) = 0.35.[6] Using this value for R(0.23), using the values R(0.56) = 0.55, and R(0.77) = 0.65, and, finally, using Eq. 5.4, we obtain the following corrected reported probabilities:

[5] For each midpoint between two given probabilities p, we calculated the average report for the adjacent probabilities. For instance, to compute the R(p) for p = 0.625, we averaged the reported probabilities for p = 0.6 and those for p = 0.65.

[6] We have 0.23 = 0.865×0.225 + 0.135×0.275, R(0.225) = 0.345, and R(0.275) = 0.382, so that R(0.23) = $R\big(0.865\times0.225 + 0.135\times0.275\big) = 0.865\times R(0.225) + 0.135\times R(0.275) = 0.865\times0.345 + 0.135\times0.382 = 0.35$.

$B(S) = R^{-1}(0.35) = 0.23$; $B(T) = R^{-1}(0.55) = 0.56$; $B(I) = R^{-1}(0.65) = 0.77$;
the additivity bias is $0.23 + 0.56 - 0.77 = 0.02$. (6.3)

The risk-correction has reduced the violation of additivity, which, according to Bayesian principles, can be interpreted as a desirable move towards rationality. In the experiment described in the following sections we will see that this effect is statistically significant for single evaluations (treatment "t=ONE"), but is not significant for repeated payments and decisions (treatment "t=ALL").

It is statistically preferable to fit data with smoother curves than those resulting from linear interpolation. We derived "decision-theoretic" parametric curves for R(p) from Corollary 5.3, with further assumptions explained in §8.1.[7] The resulting curve for participant 14 is given in Figure 6.2. The equality $B = R^{-1}(r)$ and this curve imply

$B(S) = R^{-1}(0.35) = 0.24$; $B(T) = R^{-1}(0.55) = 0.59$; $B(I) = R^{-1}(0.65) = 0.76$; the additivity bias is $0.24 + 0.59 - 0.76 = 0.07$, (6.4)

again reducing the uncorrected additivity bias. For this participant the quadratic curve, explained in §12, happens to be indistinguishable from the decision theoretic curve.

7. AN EXPERIMENTAL APPLICATION OF RISK CORRECTIONS: METHOD

The following five sections (§§7-11) present the third part of this paper. These sections give an experimental implementation of our new measurement method. We first describe the two main treatments in detail. §10 presents a third, control, treatment.

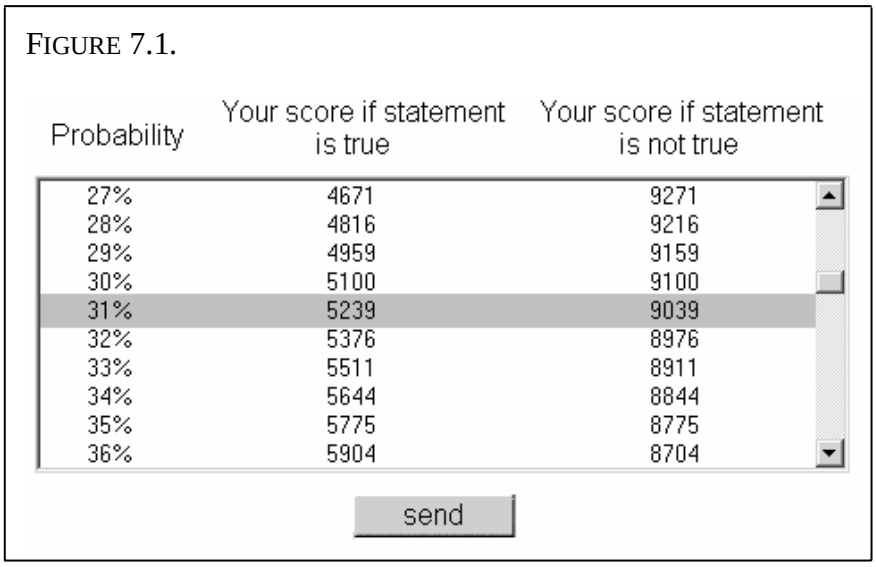
---

[7] The decision-theoretic curve in the figure is the function $p = B(E) = \dfrac{r}{r + (1-r)\dfrac{0.26\left(1-(1-r)^2\right)^{-1.26}}{0.26(1-r^2)^{-1.26}}}$, in agreement with Corollaries 5.3 and 5.4, where we estimated $w(p) = p$ and found $\rho = -0.26$ as the optimal value for $U(x)$ in Eq. 7.1.

Participants. For the first two treatments, N = 93 students from a wide range of disciplines (45 economics; 13 psychology; 35 other disciplines) participated in the experiment. They were self-selected from a mailing list of approximately 1100 people.

Procedure. Participants were seated in front of personal computers in 6 groups of approximately 16 participants each. They first received an explanation of the QSR, given in Appendix D. Then, for each uncertain event, participants could first report a probability (in percentages) by typing in an integer from 0 to 100. Subsequently, the confirmation screen displayed a list box with probabilities and the corresponding score when the event was true/not true, illustrated in Figure 7.1.

FIGURE 7.1.



All the figures (including Figure 6.1) are reproduced here in black and white; in the experiment we used colours to further clarify the figures. The entered probability and the corresponding score were preselected in this list box. The participant could confirm the decision or change to another probability by using the up or down arrow or by scrolling to another probability using the mouse. The event itself was also visible on the confirmation screen. Thus, the reported probability r finally resulted for the uncertain event.

Stimuli. The participants provided 100 reported probabilities r for events with unknown probabilities in the *stock-price part* of the experiment. For these events, we fixed June 1, 1991, as the "evaluation date." The uncertain events always concerned the question whether or not the price of a stock would lie in a target-interval seven months after the evaluation date. For each stock, the participants received a graph depicting the price of the stock on 0, 1,

2, 3, 4, and 5 months prior to the evaluation date, as well as an upper and lower bound to the price of the stock on the evaluation date. Figure 6.1 (without the braces and letters) gives an example of the layout. We used 32 different stocks, all real-world stock market data from the 1991 Amsterdam Stock Exchange. After 4 practice questions, the graph of each stock-price was displayed once in the questions 5-36, once in the questions 37-68, and once in the questions 69-100. We, thus, obtained three probabilistic judgments of the performance of each stock, once for a large target-interval and twice for the small target-intervals that partitioned the large target-interval (see Figure 6.1). We partially randomized the order of presentation of the elicitations. Each stock was presented at the same place in the first, second, and third 32-tuple of elicitations, so as to ensure that questions pertaining to the same stock were always far apart. The order of presentation of the one large and the two small intervals for each stock was not randomized stochastically, but was varied systematically, so that all orders of big and small intervals occurred equally often. We also maximized the variation of whether small intervals were both very small, or were both moderately small, or one was very small and one was moderately small.

In the *calibration part* of the experiment, participants essentially made the same decisions as in the stock-price part, but now for 20 events with objective probabilities. We used two 10-sided dice to determine the outcome of the different prospects. One die determined the first digit and the other determined the second digit of a random number below 100. An event with probability 0.25 was, for instance, described as "The outcome of the roll with two 10-sided dice is in the range 01–25." The subjects then chose r in $(1–(1–r)^2)_E(1–r^2)$ with E the event as described. This amounts to choosing the optimal r for objective probability distributions $(1–(1–r)^2)_{0.25}(1–r^2)$. We obtained measurements of the reported probabilities corresponding to the objective probabilities 0.05, 0.10, 0.15, …, 0.85, 0.90, and 0.95 (we measured the objective probability 0.95 twice). The decision screen was very similar to Figure 7.1, but we wrote "row-percentage" instead of "probability" and "your score if the roll of the die is 01-25" instead of "your score if the statement is true;" and so on.

Motivating Participants. Depending on whether the uncertain event obtained or not, and on the reported probability for the uncertain event, a number of points were determined for each question by means of the QSR (Eq. 2.1), using 10000 points as the unit of payment so as to have integer scores with four digits of precision. Thus, the maximum score for one question

was 10000, the minimum score was 0, and the certain score resulting from reported probability 0.5 was 7500 points.

In treatment t=ALL, the sum of all points for all questions was calculated for each participant and converted to money using an exchange rate of 60000 points = €1, yielding an average payment of €15.05 per participant. For the calibration part, we then used a box with 20 separate compartments containing pairs of 10-sided dice to determine the outcome of each of the 20 prospects at the same time for the treatment t=ALL.

In treatment t=ONE, the random incentive system was used. That is, at the end of the experiment, 1 out of the 120 questions that they answered was selected at random for each participant, and the points obtained for this question were converted to money using an exchange rate of 500 points = €1, yielding an average payment of €15.30 per participant.

All payments were done privately at the end of the experiment.

Analysis. For the calibration part we only need to analyze probabilities of 0.5 or higher, by Eq. 2.3 (see also Observation A.2). Every observation for $p < 0.5$ amounts to an observation for $p' = 1-p > 0.5$. It implies that we have two observations for all $p > 0.5$ (and three for $p = 0.95$).

We first analyze the data at the group level, assuming homogeneous participants. We start from general probabilistic sophistication. Note that this model can be estimated using a non-parametric procedure. If the agent is willing to go through a large series of correction questions, it is possible to measure the corresponding reported probability of each objective probability repeatedly. In this way, an accurate estimate of the whole correction curve can be obtained without making assumptions about the utility function or the weighting function. This procedure is appropriate if the goal is to correct an expert, e.g., correct the reports provided by a weatherman. In applications of experimental economics where subjects participate for a limited amount of time, the researcher will only be able to collect a limited number of observations of the correction curve. Then it is more appropriate to follow a parametric approach to elicit the curve that best fits the observations. In this paper, we used parametric fittings. For U we used the *power utility with parameter* ρ, also known as the

family of constant relative risk aversion (CRRA)[8], and the most popular parametric family for fitting utility, which is defined as follows:

For $\rho>0$: $U(x) = x^{\rho}$;
for $\rho=0$: $U(x) = ln(x)$;
for $\rho<0$: $U(x) = -x^{\rho}$. (7.1)

It is well-known that the unit of payment is immaterial for this family. The most general family that we consider for w(p) is Prelec's (1998) two-parameter family

$$w(p) = \left(exp(-\beta(-ln(p))^{\alpha})\right), \tag{7.2}$$

chosen for its analytic tractability and good empirical performance. We will mostly use the one-parameter subfamily with $\beta=1$, as in Eq. 4.1, for reasons explained later. Substituting the above functions yields

$$B(E) = \exp\left(-\left(\frac{-ln\left(\frac{r(2r-r^2)^{1-\rho}}{(1-r)(1-r^2)^{1-\rho} + r(2r-r^2)^{1-\rho}}\right)}{\beta}\right)^{1/\alpha}\right).$$

for Eq. 5.2.

The model we estimate for each subject separately is as follows:

$$R_k(j/20) = h(j/20,\alpha,\rho) + \varepsilon_k(j/20). \tag{7.3}$$

Here $R_k(j/20)$ is the reported probability of the participant for known probability $p=j/20$ ($10 \le j \le 19$) in treatment t (t = ALL or t = ONE) for the k^{th} measurement for this probability, with only k=1 for j = 10, k = 1,2 for $11 \le k \le 18$, and k = 1,2,3 for j = 19. With β set equal to 1, α is the remaining probability-weighting parameter (Eq. 7.2), and ρ is the power of utility (Eq. 7.1). The function h is the inverse of Eq. 5.3. Although we have no analytic expression for this inverse, we could calculate it numerically in the analyses. The error terms $\varepsilon_k(j/20)$ are drawn from a truncated normal distribution with mean 0 and variance σ^2. The distribution of the error terms is truncated because reported probabilities below 0 and above 1 are excluded

[8] We avoid the latter term because, in nonexpected utility models as relevant for this paper, risk aversion depends not only on the curvature of utility.

by design. Error terms are identically and independently distributed across choices. We employed maximum likelihood to estimate the parameters of Eq. 7.3.

We also carried out an analysis at the aggregate level of the calibration part, with α_t and ρ_t, i.e. with these parameters depending on the treatment but not on the participant. To correct for individual differences, we added an individual-specific constant $c_{s,t}$ to the equation where s refers to the participant and t to the treatment:

$$R_{s,t,k}(j/20) = h(j/20,\alpha_t, \rho_t) + c_{s,t} + \varepsilon_{s,t,k}(j/20,\sigma_t^2)\,. \qquad (7.4)$$

Here the error terms are independent across subjects, treatments, and choices.

In the stock-price part, violations of additivity were tested. With I the large interval of a stock, being the union $S \cup T$ of the two small intervals S and T, additivity of the uncorrected reported probabilities implies

$$r_S + r_T - r_I = 0 \qquad (7.5)$$

Hence, $r_S + r_T - r_I$ is an index of deviation from additivity, which we call the *additivity bias* of r.

Under the null hypothesis of additivity for corrected reported probabilities B, binary additivity holds, and we can obtain $B(S) = 1 - B(S^c)$ for small intervals S in the experiment (cf. Eq. 2.2). Thus, under additivity of B, we have

$$B(S) + B(T) - B(I) = 0. \qquad (7.6)$$

Hence, $B(S) + B(T) - B(I)$ is an index of deviation from additivity of B, and is B's *additivity bias*.

We next discuss tests of the additivity bias. For each individual stock, and also for the average over all stocks, we tested for both treatments t=ONE and t=ALL: (a) whether the additivity bias was zero or not, both with and without risk correction; (b) whether the average additivity bias, as relevant for aggregated group behaviour and expert opinions, was enlarged or reduced by correction; (c) whether the absolute value of the additivity bias, as relevant for additivity at the individual level, was enlarged or reduced by correction. We report only the tests for averages over all stocks.

8. RESULTS OF THE CALIBRATION PART

Risk-corrections and, in general, QSR measurements, do not make sense for participants who are hardly responsive to probabilities, so that R(p) is almost flat on its entire domain. Hence, we kept only those participants for whom the correlation between reported probability and objective probability exceeded 0.35. We thus dropped 4 participants. The following analyses are based on the remaining 89 participants.

8.1. *Group Averages*

We did several tests using Eq. 7.2 with β as a free (treatment-dependent or -independent) variable, but β's estimates added little extra explanatory power to the other parameters and usually were close to 1. Hence, we chose to focus on a more parsimonious model in which the restriction $\beta_{ONE} = \beta_{ALL} = 1$ is employed. Table 8.1 lists the estimates for the model of Eq. 7.4 for β=1 (Eq. 4.1 instead of Eq. 7.2), together with the estimates of some models with additional restrictions. We first give results at the aggregate level. Because there turns out to be a strong correlation between the α and the ρ parameters, estimation results where both parameters are estimated simultaneously cannot be trusted, and we only report the results where either α or ρ is estimated.

TABLE 8.1. Estimation results at the aggregate level

Row	Restrictions	σ_{ONE}	α_{ONE}	ρ_{ONE}	σ_{ALL}	α_{ALL}	ρ_{ALL}	–LogL
1	$\alpha_{ONE} = \alpha_{ALL} = \rho_{ONE} = \rho_{ALL} = 1$	9.00** (0.21)	–	–	8.36** (0.20)	–	–	6373.21
2	$\alpha_{ONE} = \alpha_{ALL} = 1$	8.73** (0.20)	–	0.43** (0.09)	8.36** (0.20)	–	0.94** (0.07)	6345.43
3	$\rho_{ONE} = \rho_{ALL} = 1$	8.82** (0.21)	0.69** (0.03)	–	8.35** (0.20)	1.09** (0.07)	–	6354.14
4	$\alpha_{ONE} = \alpha_{ALL} = \rho_{ONE} = 1$	9.00** (0.21)	–	–	8.36** (0.20)	–	0.94** (0.07)	6372.87
5	$\alpha_{ONE} = \alpha_{ALL} = \rho_{ALL} = 1$	8.73** (0.20)	–	0.43** (0.09)	8.36** (0.20)	–	–	6345.77
6	$\alpha_{ONE} = \alpha_{ALL} = 1$, $\rho_{ONE} = \rho_{ALL}$	8.78** (0.21)	–	0.70** (0.06)	8.41** (0.20)	–	–	6556.48

Standard errors in parentheses, ** (*) denotes significance at the 1% (5%) level.

Overall Need for Risk-Correction. The 1st row of Table 8.1 shows the results without any correction. The 2nd row presents the results when utility curvature is introduced. The likelihood improves significantly (Likelihood Ratio test, p = 0.01) and substantially, so that risk-correction is called for. Risk-correction can also be done by probability weighting. This is done in the 3rd row of the table. Probability weighting also increases the likelihood of observing the data significantly compared with the model without correction, but less so than utility curvature does. Therefore, in the remainder of the paper we focus on risk-correction obtained through utility curvature.

Comparing the Two Treatments. At the aggregate level, risk-correction is needed less in treatment ALL than in treatment ONE, as the 4th and 5th rows show. In treatment ONE the likelihood is improved significantly (compare the 5th and the 1st row; Likelihood Ratio test, p = 0.01) but in treatment ALL the likelihood is not improved significantly (compare the 4th and the 1st row: Likelihood Ratio test, p > 0.10). We obtain $\rho_{ONE} < \rho_{ALL}$: if only one decision is paid out, then participants exhibit more concave curvature of utility than when all decisions are paid out. Given the same degree of probability weighting, it implies more risk aversion for t=ONE than for t=ALL (and R closer to 0.5). The finding is supported by comparing the 6th row of Table 8.1 with the restriction $\rho_{ONE} = \rho_{ALL}$, to the 2nd row. This restriction significantly reduces the likelihood of observing the data (Likelihood Ratio test, p = 0.01).

FIGURE 8.1. Corrected versus uncorrected probability

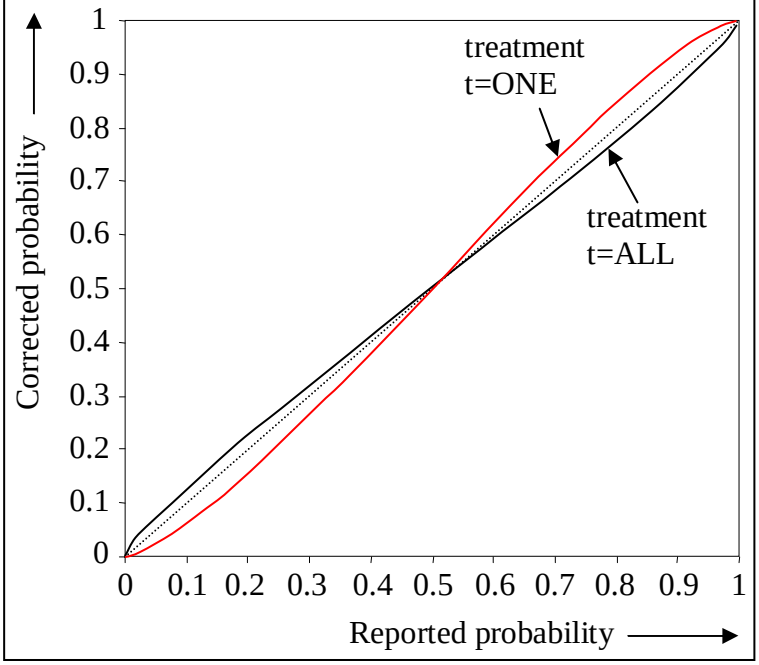
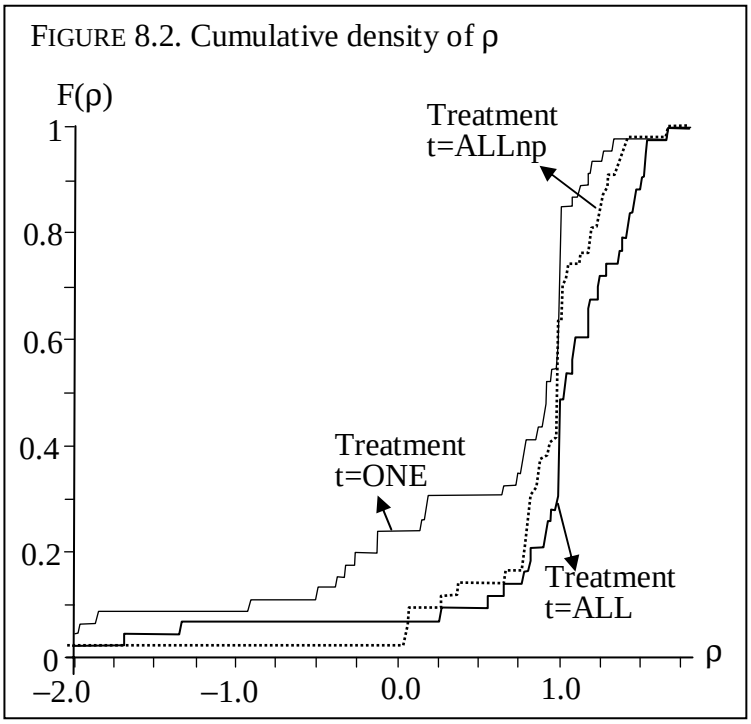


Figure 8.1, based on the estimates reported in the 2nd row of Table 8.1, displays the resulting average risk-correction for the two treatments separately. The figure illustrates that risk correction is clearly needed at the aggregate level in treatment ONE.

8.2. *Individual Analyses*

Need for Risk-Correction at the Individual Level. There is considerable heterogeneity in each treatment. Whereas the corrections required were small at the level of group averages, they are big at the individual level. This appears from Figure 8.2, which displays the cumulative distribution of the (per-subject) estimated ρ-coefficients for each treatment, assuming $\alpha = \beta = 1$. (The figure also displays a treatment t=ALLnp that will be explained in Section 10.) There are wide deviations from the value ρ=1 (i.e. no correction) on both sides. As seen from the group-average analysis, there are more deviations at the risk-averse side of $\rho < 1$.

FIGURE 8.2. Cumulative density of ρ



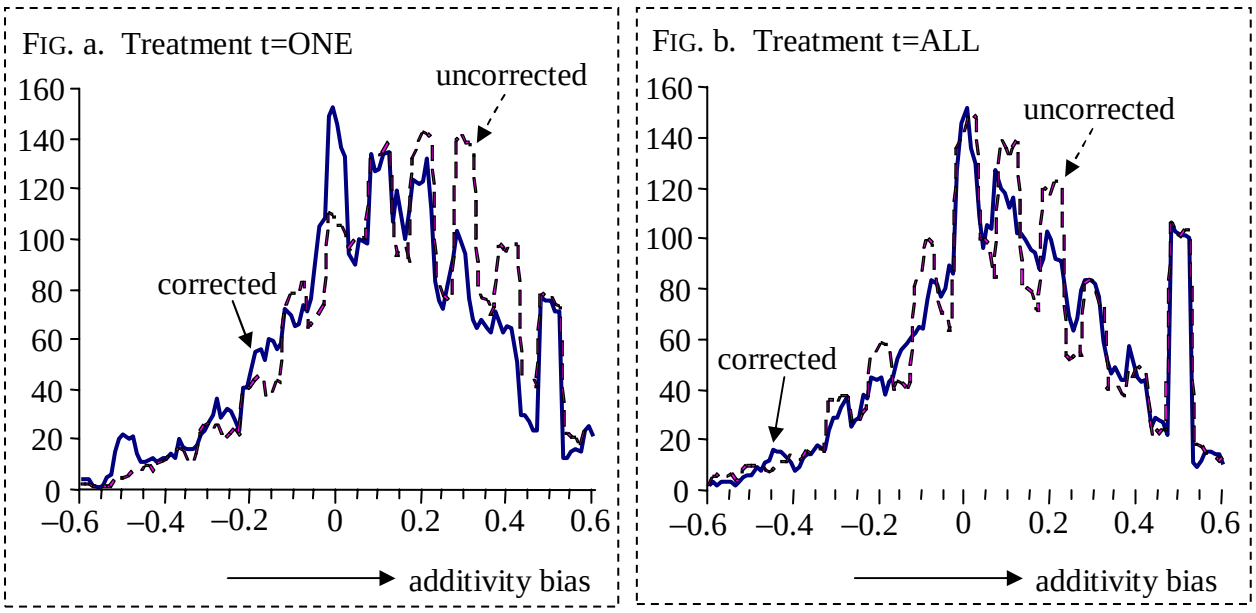
Comparing the Two Treatments. The ρ-coefficient distribution of treatment t=ONE dominates the ρ-coefficient distribution of treatment t=ALL. Thus, the ρ-coefficients for

t=ONE are lower than for t=ALL (p=0.001: two-sided Mann-Whitney test). It confirms the result from Table 8.1 that there is more risk aversion for group averages, moving R in the direction of 0.5, for t=ONE than for t=ALL. The figure also shows that, in an absolute sense, there is more deviation from ρ=1 for t=ONE than for t=ALL, implying that there are more deviations from expected value and more risk corrections for t=ONE than for t=ALL.

Unlike the median ρ-coefficients that are fairly close to each other for the two treatments (0.92 for t=ONE versus 1.04 for t=ALL), the mean ρ-coefficients are substantially different (0.24 for t=ONE versus 0.91 for t=ALL), which is caused by skewedness to the left for t=ONE. That is, there is a relatively high number of strongly risk-averse participants for t=ONE. Analyses of the individual ρ parameters (two-sided Wilcoxon signed rank sum tests) confirm findings of group-average analyses in the sense that the ρ-coefficients are significantly smaller than 1 for t=ONE ($z = -3.50$, $p = 0.0005$), but not for t=ALL ($z = 1.42$, $p = 0.16$).

9. RESULTS FOR THE STOCK-PRICE PART: RISK-CORRECTION AND ADDITIVITY

FIGURE 9.1. Empirical density of additivity bias for the two treatments



For each interval $[\frac{j-2.5}{100}, \frac{j+2.5}{100}]$ of length 0.05 around $\frac{j}{100}$, we counted the number of additivity biases in the interval, aggregated over 32 stocks and 89 individuals, for both treatments. With risk-correction, there were 65 additivity biases between 0.375 and 0.425 in the treatment t=ONE, and without risk-correction there were 95 such; and so on.

All comparisons in this section are based on two-sided Wilcoxon signed rank sum tests. Figure 9.1 displays the data, aggregated over both stocks and individuals, of the additivity biases for t=ONE and for t=ALL. The figures show that the additivity bias is more often positive than negative, in agreement with common findings in the literature (Tversky and Koehler 1994). Indeed, for virtually all stocks the additivity bias is significantly positive for both treatments, showing in particular that additivity does not hold. This also holds when taking the average additivity bias over all stocks as one data point per participant ($z = 5.27$, $p < 0.001$ for t=ONE, $z = 4.35$, $p < 0.001$ for t=ALL). We next consider whether risk corrections reduce the violations of additivity.

Let us first consider treatment t=ONE. Here the risk corrections reduce the average additivity bias significantly for 27 of the 32 stocks, and enlarge it for none. We only report the statistics for the average additivity bias over all stocks taken as one data point per participant, which has overall averages 0.163 (uncorrected) and 0.120 (corrected), with the latter significantly smaller ($z = 3.21$, $p = 0.001$). For assessing the degree of irrationality (additivity-violation) at the individual level, the absolute values of the additivity bias are relevant. For t=ONE, Figure 9.1 suggests that these are smaller after correction, because on average the corrected curve is closer to 0 on the x-axis. These absolute values were significantly reduced for 9 stocks, and enlarged for none. Again, we only report the statistics for the average absolute value of the additivity bias over all stocks taken as one data point per participant, which has overall averages 0.239 (uncorrected) and 0.228 (corrected), with the latter significantly smaller ($z = 2.26$, $p = 0.02$).

For t=ALL, risk corrections did not significantly alter the average additivity bias. More specifically, it gave a significant increase for 3 stocks and a significant decrease for 1 stock, which, for 32 stocks, suggests no systematic effect. The latter was confirmed when we took the average additivity bias over all stocks for each individual, with no significant differences generated by correction (average 0.128 uncorrected and average 0.136 corrected; $z = -1.64$, $p = 0.1$). Similar results hold for absolute values of additivity biases, which gave a significant increase for 1 stock and a significant decrease for no stocks. Taking the average additivity bias over all stocks as one data point per participant (average 0.237 uncorrected and average 0.239 corrected; $z = -0.36$, $p = 0.70$) also gave no significant difference.

Risk correction reduces the additivity bias for treatment t=ONE to a level similar to that observed for t=ALL (averages 0.120 and 0.136). The overall pattern is that beliefs for t=ONE after correction, and for t=ALL both before and after correction, exhibit a similar degree of violation of additivity, which is clearly different from zero. The additivity bias is not completely caused by nonlinear risk attitudes when participants report probabilities, but has a genuine basis in beliefs.

10. A TREATMENT WITHOUT EXPLICIT REFERENCE TO BELIEFS OR PROBABILITY

This section briefly reports the results of a robustness check of our experimental design. In agreement with the current practice of scoring rules, the instructions of our main treatments repeatedly used the terms probability and belief. These terms may have influenced the subjects. To assess such influences, we performed a control treatment in which we did not refer to probabilities or beliefs.[9] Thus, in Figure 7.1 we now used the expression "choose a number" instead of probability. In the instructions of this control treatment, we similarly asked subjects to choose numbers without calling them probabilities, and we dropped all interpretations of likelihood. In this manner we ran the control treatment for t=ALL. We chose t=ALL rather than t=ONE because the former is most commonly used in applications of proper scoring rules. We refer to the control treatment as t=ALLnp (np for no probabilities). N=44 students participated. The number of participants dropped from the analysis because their correlation between reported and objective probability was below 0.35, was now 2. In all other respects the new treatment was identical to the t=ALL treatment in the main experiment.

The results confirmed all patterns and inequalities found for t=ALL. We give some numerical details for individual analyses. The ρ's of t=ALLnp are not significantly different from those of t=ALL ($z = 1.57$, $p=0.12$), with a similar median (1.00 for t=ALLnp versus 1.04 for t=ALL) and mean (0.80 for t=ALLnp versus 0.91 for t=ALL). They, accordingly, are not significantly below 1 either ($z = 0.52$, $p = 0.60$), and they also exceed the ρ for t=ONE ($z = -2.30$, $p = 0.02$).

[9] This treatment was recommended to us by a referee and the editor.

The additivity bias is, again, positive, showing that additivity is violated, for most individual stocks. It also is when taking the average additivity bias over all stocks per participant ($z = 4.47$, $p < 0.001$). Risk corrections did not significantly increase or decrease the average additivity bias for any stock. For the (absolute) average additivity bias over all stocks per participant, we again found no significant difference between the non-corrected and corrected average additivity bias ($z = 0.378$, $p = 0.71$; $z = 0.265$, $p = 0.79$ for absolute values). The risk-corrected average additivity bias for t=ALLnp being virtually the same as for t=ALL (0.126 versus 0.136) obviously implies that it is also equal to the one for t=ONE (0.120). In summary, all the results for the t=ALL treatment were confirmed by the t=ALLnp treatment, suggesting that the explicit use of the term probability in our instructions did not alter the results.

11. DISCUSSION OF EXPERIMENT

Methods. We chose the evaluation date (June 1, 1991) sufficiently long ago to ensure that participants would be unlikely to recognize the stocks or have private information about them. In addition, no numbers were displayed on the vertical axis, making it extra hard for participants to recognize specific stocks. We, thus, ensured that participants based their probability judgments entirely on the prior information about past performance of the stocks given by us. Given the large number of questions, it is unlikely that participants noticed that the graphs were presented more than once (three times) for each stock. Indeed, in informal discussions after the experiment no participant showed awareness of this point.

In some studies in the literature, the properness of scoring rules is explained to participants by stating that it is in their best interest to state their true beliefs, either without further explanation, or with the claim added that they will thus maximize their "expected" money. A drawback of this explanation is that expected value maximization is empirically violated, which is the central topic of this paper (§3), so that the recommendation is debatable. We, therefore, used an alternative explanation that relates properness of one-off events to observed frequencies of repeated events (Appendix D).

Optimal Incentive Scheme. After some theoretical debates about the random incentive system (Holt 1986), as in our treatment t=ONE, the system was tested empirically and found

to be incentive compatible (Lee 2008; Starmer and Sugden 1991). It is today the almost exclusively used incentive system for the measurement of individual preferences (Holt and Laury 2002; Harrison et al. 2002; Myagkov and Plott 1997). Unlike repeated payments it avoids income effects such as Thaler and Johnson's (1990) house money effect, and the drift towards expected value and linear utility that is commonly generated by repeated choice.[10] For the purpose of measuring individual preference, the treatment t=ONE is, therefore, preferable. When the purpose is, however, to derive subjective probabilities from proper scoring rules, and no risk-correction is possible, then a drift towards expected value is actually an advantage, because uncorrected proper scoring rules assume expected value. This point agrees with our findings, where less risk-correction was required for the t=ALL treatment. Li (2007) discussed other arguments for and against repeated rewarding when events are not verifiable and binary rewards have to be used.

For some applications group averages of probability estimates are most relevant, such as when aggregating expert judgments or predicting group behaviour. Then our statistical results regarding "non-absolute" values of reported probabilities are most relevant. For the assessment of rationality at the individual level, absolute values of the additivity biases are most relevant.

Choice of Parameters. The lack of extra explanatory power of parameter β in Eq. 7.2 should come as no surprise because β and α imply similar phenomena on [0.5,1], increasing risk aversion there. They mainly deviate from one another on [0,0.5], where β continues to enhance risk aversion but α enhances the inverse-S shape that is mostly found empirically. The domain [0,0.5] is, however, not relevant to our study (Observation A.2).

Pragmatic Applications. More tractable families can be used to fit the reported probabilities than the decision-theory-based curves that we used. For example, in Figure 6.2 we also used quadratic regression to find the curve $p = a + br + cr^2$ that best fits the data. For most participants, the curve is virtually indistinguishable from the decision-theoretic curve. This observation, together with Corollary 5.4 which demonstrates that we only need the readily observable reported probabilities and not the actual utility function or probability weighting function to apply our method, shows that applications of our method are straightforward. The

[10] It is required that the repeated choices are perceived as sufficiently uncorrelated. Correlation can enhance the perception of, and aversion to, ambiguity (Halevy and Feltkamp 2005).

theoretical analysis of this paper, and the decision-theory based curve-fitting that we adopted, served to prove that our method is in agreement with modern decision theories. If this thesis is accepted, and the only goal is to obtain corrected reported probabilities, then one may choose the pragmatic shortcuts just described.

General Discussion. We emphasize that the biases due to violations of expected value that we correct for need not concern mistakes or irrationalities in decision making. Deviations from risk neutrality need not be irrational and, according to some, even deviations from Bayesian beliefs need not be irrational, nor need the corresponding ambiguity attitudes be irrational (Gilboa and Schmeidler 1989; Schmeidler 1989). The required corrections concern empirical deficiencies of the model of expected value, i.e. they concern biases on the part of the researchers analyzing the data.

Under proper scoring rules, beliefs are derived solely from decisions, and Eq. 2.1 is taken purely as a decision problem, where the only goal of the agent is to optimize the prospect received. We considered two treatments that explicitly referred to probabilities and beliefs, and a treatment that did not do so, finding no differences between the two. Thus, this paper has analyzed proper scoring rules purely from the decision-theoretic perspective supported with real incentives, and has corrected only for biases resulting therefrom. Many studies have investigated direct judgments of belief without real incentives, and then many other aspects play a role, leading for instance to the often found overconfidence. Such introspective effects are beyond the scope of this paper.

An immediate advantage of our calibration measurement, prior to any theoretical analysis, is that it helps to identify subjects whose understanding of the concepts to be measured is below what is minimally acceptable. Indeed, subjects for whom the correlation between objective and reported probabilities is very low clearly have little clue what likelihood means. Their reported probabilities are of so little interest that we recommend dropping them from the sample. If we are interested in the beliefs of such subjects in more elaborate studies, then further teaching and learning will be called for.

The experimental data show that a substantial correction of reported probabilities needs to be made for a subset of the subjects. The fraction of the sample that needs substantial corrections is larger when only a single large-stake decision is paid than when repeated small decisions are paid. Our conclusion is that it is desirable to correct agents' reported probabilities elicited with scoring rules, especially if only a single large-stake decision is paid. A drawback of our method is, obviously, that it requires individual measurements of

QSRs for given probabilities. If it is not possible to obtain individual measurements of the correction curve, then it will be useful to use best-guess corrections: for instance, through averages obtained from individuals as similar as possible. Thus, at least, the systematic error for the group average to risk attitude has been corrected for as well as is possible without requiring extra measurements. In this respect the average curves in Figure 8.1 are reassuring for existing studies, because these curves suggest that only small corrections were needed for the group averages in our context.

Several methods have been used in the literature to measure the subjective degree of belief of an agent in an event E. Mostly these have been derived from: (a) binary preferences, which only give inequalities or approximations; (b) binary indifferences, which are hard to elicit, e.g. by the complex Becker-DeGroot-Marschak mechanism (Braga and Starmer 2005; Karni and Safra 1987) or bisection (Abdellaoui, Vossman, and Weber (2005); (c) introspection, which is not revealed preference based let alone incentive compatible. Proper scoring rules provide an efficient way to measure subjective beliefs while avoiding the problems mentioned.

12. THEORETICAL DISCUSSION

A way to reveal B(E) from observed choice, as an alternative to our method, is by revealing the *matching probability* p of event E, defined through the equivalence

$$x_p y \sim x_E y \qquad (12.1)$$

for some preset $x > y$, say $x = 100$ and $y = 0$. Then $w(B(E))(U(x)-U(y)) = w(p)(U(x)-U(y))$, and $B(E) = p$ follows. Wakker (2004) discussed the interpretation of Eqs. 5.1 and 12.1 as belief. Matching probabilities were commonly used in early decision analysis (Raiffa 1968, §5.3; Yates 1990, pp. 25-27) under the assumption of expected utility. A recent experimental measurement is in Holt (2006, Ch. 30), who also assumed expected utility. Abdellaoui, Vossman, and Weber (2005) measured and analyzed them in terms of prospect theory, as does our paper. A practical difficulty is that the measurement of matching probabilities requires the measurement of indifferences, and these are not easily inferred from choice. For example, Holt (2006) used the Becker-deGroot-Marschak mechanism, and Abdellaoui, Vossman, and Weber (2005) used a bisection method. If we want to measure B(E) for only

few events E, then matching probabilities provide a tractable alternative to our correction method. Our method is more efficient when dealing with many events. The measurement of the correction curve is a one-time investment that can next be applied to an unlimited number of events.

Another way to correct reported probabilities is through calibration. Then many reported probabilities are collected over time and are related to observed relative frequencies. Calibration has been studied in game theory (Sandroni, Smorodinsky, and Vohra 2003), and has been applied to weather forecasters (Murphy and Winkler 1974). It needs extensive data, which is especially difficult to obtain for rare events such as earthquakes, and further assumptions such as the stability of the characteristics of these events over time. Clemen and Lichtendahl (2005) discussed these drawbacks and proposed correction techniques for probability estimates in the spirit of our paper, but still based these on traditional calibration techniques. Our correction ("calibration") technique is considerably more efficient than the traditional ones. It shares with Prelec's (2004) method the advantage that we need not wait until the truth or untruth of uncertain events has been revealed in order to implement it.

Allen (1987) proposed to avoid biases of the QSR due to nonlinear utility by paying in terms of the probability of winning a prize instead of in terms of money, and this procedure was implemented by McKelvey and Page (1990). The procedure, however, only works if expected utility holds, and there is much evidence against this assumption. Indeed, Selten, Sadrieh, and Abbink (1999) showed empirically that payment in probability generates more deviations from risk neutrality than payment in money.

The decision-based distortion in the direction of 0.5 due to risk aversion in §4 is opposite to the overconfidence (probability judgments too far from 0.5) mostly found in direct judgments of probability without real incentives (McClelland and Bolger 1994), and found among experts seeking to distinguish themselves (Keren 1991, p. 224 and 252; the "expert bias", Clemen and Rolle 2001). Similar optimistic and pessimistic distortions of probability can result from nonlinear utility if the probability considered is a consensus probability for a group of individuals with heterogeneous beliefs (Jouini and Napp 2007).

The curve for nonEU in Figure 4.1 is flat around $p = 0.5$, or more precisely, on the probability interval [0.43, 0.57]. For probabilities from this interval the risk aversion generated by nonexpected utility is so strong that the agent goes for maximal safety and chooses $r = 0.5$, corresponding with the sure outcome 0.75 (cf. Manski 2004, footnote 10; Segal and Spivak 1990). Such a degree of risk aversion is not possible under expected utility, where $r = 0.5$ can happen only for $p = 0.5$ (Observation 3.3). This observation cautions

against assigning specific levels of belief to observations r = 0.5, because proper scoring rules may be insensitive to small changes in the neighbourhood of p = 0.5. It in fact means that there the scoring rules, traditionally called proper, are not really proper there.

B captures the component of decision making separate from risk attitude. It is common in decision theory to interpret factors separate from risk attitude as ambiguity. Then B reflects ambiguity attitude. There is no consensus about the extent to which ambiguity reflects non-Bayesian beliefs, and to what extent it reflects non-Bayesian decision attitudes separate from belief. If the equality $B(E) + B(E^c) = 1$ (*binary additivity*) is violated, then it can further be debated whether $B(E)$ or $1 - B(E^c)$ is to be taken as an index of belief (or of ambiguity). Such interpretations have not yet been settled, and further studies are called for. We have mostly referred to B as reflecting beliefs, in order to stay as close as possible to the terminology used today in the literature on proper scoring rules. Irrespective of the interpretation of B, it is clear that the behavioural component of risk attitude should be filtered out before an interpretation of belief can be considered. This paper shows how this filtering out can be done. In Schmeidler (1989), the main paper initiating Eqs. 3.1 and 3.2, w was assumed to be linear, with expected utility for given probabilities, and W coincided with B. Schmeidler interpreted this component as reflecting beliefs.

As is common in the mechanism design literature, our correction procedure assumed deterministic choice. A fundamental question concerns how the mechanism performs when agents make mistakes, as in the random utility model (Luce 1959; McFadden 1974, 1976). Such mistakes will affect the optimal elicitation procedure. These issues are relevant to the entire mechanism design literature and are a topic for future research.

13. CONCLUSION

This paper has applied modern theories of risk and ambiguity to proper scoring rules. Mutual benefits have resulted for users of proper scoring rules and for analysts of risk and ambiguity. For the former we have shown which distortions affect their common measurements and how large these distortions are, using theories that are descriptively better than the expected value hypothesis still common today in applications of proper scoring rules. We have provided a procedure to correct for the aforementioned distortions, and a theoretical foundation has been given for interpretations of the resulting measurements as (possibly non-

Bayesian) beliefs and ambiguity attitudes. For analyses of risk and ambiguity we have shown how the remarkable efficiency of proper scoring rules can be used to measure and analyze subjective beliefs and ambiguity attitudes in ways more tractable than is possible with the binary preferences traditionally used.

The feasibility and tractability of our method have been demonstrated in an experiment, where we used it to investigate some properties of beliefs and quadratic proper scoring rules. We found, for instance, that our correction method reduces the violations of additivity in subjective beliefs but does not eliminate them. It confirms that beliefs are genuinely non-Bayesian and that ambiguity attitudes play a central role in proper scoring rules.

APPENDIX A. TECHNICAL REMARKS

For QSR-prospects in Eq. 2.1, every choice $r < 0$ is inferior to $r = 0$, and $r > 1$ is inferior to $r = 1$. The optimization problem does not change if we allow all real r, instead of $0 \le r \le 1$. Hence, solutions $r = 0$ or $r = 1$ can be treated as interior solutions, and they satisfy the first-order optimality conditions.

In general, it may not be possible to derive both w and U from R(p) without further assumptions, i.e. U and w may be nonidentifiable for proper scoring rules. Under regular assumptions about U and w, however, they have some different implications. The main difference is that, if we assume that U is differentiable (as done throughout this paper) and concave, then a flat part of R(p) around 0.5 must be caused by w (Observation 3.3).

We next discuss dualities between B(E) and $1 - B(E^c)$ in more detail. Event A is (*revealed*) *more likely than* event B if, for some positive outcome x, say x = 100, the agent prefers x_A0 to x_B0. This observation is independent of the outcome $x>0$. In view of the symmetry of QSRs in Observation 2.1, for $r \neq 0.5$ the agent will always allocate the highest payment to the most likely of E and E^c. It leads to the following restriction of QSRs.

OBSERVATION A.1. Under the QSR in Eq. 2.1, the highest outcome is always associated with the most likely event of E and E^c. □

Hence, QSRs do not give observations about most likely events when endowed with the worst outcome. Similar restrictions apply to all other proper scoring rules considered in the literature so far. It implies the following result.

OBSERVATION A.2. For the QSR, only the restriction of w to [0.5,1] plays a role, and w's behavior on [0,0.5) is irrelevant. □

For our risk-corrections, we need w only on [0.5,1]. An advantage is that the empirical findings about w are uncontroversial on this domain, the general finding being that w underweights probabilities there. This holds both for the mostly found inverse-S shape (Abdellaoui 2000; Bleichrodt and Pinto 2000; Gonzalez and Wu 1999; Tversky and Kahneman 1992), and for the also often found convex shapes (Goeree, Holt, and Palfrey 2002; van de Kuilen, Wakker, and Zou 2008).[11]

Some details on weak inequalities and corner solutions are as follows. A choice of r = 0.5 may be driven by risk aversion, so that no likelihood ordering between E and E^c can then be concluded. A choice of $r \neq 0.5$ (if close to 0.5) may be driven by risk seeking with equal likelihood of E and E^c. Only interior solutions with a strict inequality $r > 0.5$ exclude that E be strictly less likely than E^c.

As with the weighting function w under risk, B is also applied only to the most likely one of E and E^c in the preceding equations, reflecting again the restriction of the QSR of Observation A.1. Hence, under traditional QSR measurements we cannot test binary additivity directly because we measure B(E) only when E is more likely than E^c. These problems can easily be amended by modifications of the QSR. For instance, we can consider prospects

$$(2-(1-r)^2)_E(1-r^2), \tag{A.1}$$

i.e. QSR-prospects as in Eq. 2.1 but with a unit payment added under event E. The classical proper-scoring-rule properties of §2 are not affected by this modification, and the results of §3 are easily adapted. With this modification, we have the liberty to combine event E with the highest outcome both if E is more likely than E^c and if E is less likely, and we avoid the

[11] On [0,0.5) the patterns is less clear, with both underweighting and overweighting (Abdellaoui 2000; Bleichrodt and Pinto 2000; Gonzalez and Wu 1999).

restriction of Observation A.1. We can then observe w of the preceding subsection, and W(E) and B(E) over their entire domain. Similarly, with prospects

$$(1-(1-r)^2)_E(2-r^2), \tag{A.2}$$

we can measure the duals $1 - W(E^c)$, $1 - w(1-p)$, and $1 - B(E^c)$ over their entire domain. In this study we confined our attention to the QSRs of Eq. 2.1 as they are classically applied throughout the literature. We have revealed their biases according to the current state of the art of decision theory, have suggested remedies whenever possible, and have signalled the problems that remain. Further investigations of the, we think promising, modifications of QSRs in the preceding equations are left to future studies.

The restrictions of the classical QSRs also hold for the experiment in this paper. There an application of the QSR to events E less likely than their complements are to be interpreted formally as the measurement of $1 - B(E^c)$. The restrictions also explain why the theorems concerned only the case of $r > 0.5$ (with $r = 0.5$ as a boundary solution).

APPENDIX B. PROOFS

For QSR-prospects in Eq. 2.1, every choice $r < 0$ is inferior to $r = 0$, and $r > 1$ is inferior to $r = 1$. The optimization problem does not change if we allow all real r, instead of $0 \le r \le 1$. Hence, solutions $r = 0$ or $r = 1$ can be treated as interior solutions, and they satisfy the first-order optimality conditions.

PROOF OF THEOREM 3.1. We write π for the decision weight W(E), and consider the general prospects $(a-b(1-r)^2)_E(a-br^2)$ for any $b > 0$ and $a \in \mathbb{R}$. Theorem 3.1 concerns the special case of $a = b = 1$. For optimality of interior solutions r, the first-order optimality condition for Eq. 3.1 is that

$$\pi U'(a-b(1-r)^2)2b(1-r) - (1-\pi)U'(a-br^2)2br = 0,$$

implying

$$\pi(1-r)U'(a-b(1-r)^2) = (1-\pi)rU'(a-br^2) \tag{B.1}$$

or

$$\pi U'(a-b(1-r)^2) = r \times \big(\pi U'(a-b(1-r)^2) + (1-\pi)U'(a-br^2)\big),$$

and Eq. 3.3 follows.

□

PROOF OF OBSERVATION 3.3. If r = 0.5 then the marginal-utility ratio in Eq. 3.3 is 1, and p = 0.5 follows. For the reversed implication, assume risk aversion. Then r > 0.5 is not possible for p = 0.5 because then the marginal-utility ratio in Eq. 3.3 would be at least 1 so that the right-hand side of Eq. 3.3 (where now W(E) = 0.5) would at most be 0.5, contradicting r > 0.5. Applying this finding to E^c and using Eq. 2.2, r < 0.5 is not possible either, and r = 0.5 follows.

Under strong risk seeking, r may differ from 0.5 for p = 0.5. For example, if $U(x) = e^{2.5x}$, then r = 0.14 and r = 0.86 are optimal, and r = 0.5 is a local infimum, as calculations can show. As an aside, the same optimal values of r result under nonexpected utility with linear U, and with w(0.5) = 0.86. Such large w-values also generate risk seeking. □

PROOF OF COROLLARY 5.2. Let r > 0.5 be optimal, and write $\pi = W(E)$. Then Eq. B.1 implies $\pi \times \big((1-r)U'(a-b(1-r)^2) + rU'(a-br^2)\big) = rU'(a-br^2)$, implying

$$\pi = \frac{r}{r + (1-r)\dfrac{U'(a-b(1-r)^2)}{U'(a-br^2)}} . \qquad \text{(B.2)}$$

Applying w^{-1} to both sides yields the theorem. □

In measurements of belief one first observes r, and then derives B(E) from it. Corollary 5.2 gave an explicit expression. In general, it does not seem to be possible to write r as an explicit expression of B(E) or, in the case of objective probabilities with B(E) = p, of the probability p.

PROOF OF COROLLARY 5.4. Theorem 3.1 implies that the right-hand side of Eq. 3.3 is r both as it is, and with p substituted for B(E) (in W(E) = w(B(E))). Because Eq. 3.3 is strictly increasing in w(B(E)), and w is strictly increasing too, p = B(E) follows. □

APPENDIX C. MODELS FOR DECISION UNDER RISK AND UNCERTAINTY

For binary (two-outcome) prospects with both outcomes nonnegative, as considered in QSRs, Eqs. 3.1 and 3.2 have appeared many times in the literature. References include the early Allais (1953, Eq. 19.1) and Edwards (1954, Figure 3) for risk and, more recently, Luce (1991) for uncertainty. The convenient feature that binary prospects suffice to identify utility U and the nonadditive $w \circ B = W$ was pointed out by Ghirardato and Marinacci (2001), Gonzalez and Wu (2003), Luce (1991, 2000), Miyamoto (1988), Pfanzagl (1959, p. 287), and Wakker and Deneffe (1996, p. 1143 and pp.1144-1145).

The convenient feature that most decision theories agree on the evaluation of binary prospects was pointed out by Miyamoto (1988), calling Eqs. 3.1 and 3.2 generic utility, and Luce (1991), calling these equations binary rank-dependent utility. It was most clearly analyzed by Ghirardato and Marinacci (2001), who called the equations the biseparable model. These three works also axiomatized the model. The agreement for binary prospects was also central in many works by Luce (e.g. Luce, 2000, Ch. 3) and in Gonzalez and Wu (2003). Only for more than two outcomes, do the theories diverge (Mosteller and Nogee 1951, p. 398; Luce 2000, Introductions to Chs. 3 and 5). Theories that also deviate for two outcomes include betweenness models (Chew and Tan 2005), the variational model (Maccheroni, Marinacci, and Rustichini 2006), and models with underlying multistage decompositions (Halevy and Feltkamp 2005; Halevy and Ozdenoren 2007; Klibanoff, Marinacchi, and Mukerji 2005; Nau 2006; Olszewski 2007).

We next describe some of the agreeing decision theories. Because we consider only nonnegative outcomes, losses play no role, and we describe prospect theory only for gains.

We begin with decision under risk, with known objective probabilities P(E). Expected utility (von Neumann and Morgenstern 1944) is the special case where w is the identity and $B(E) = P(E)$. Kahneman and Tversky's (1979) original prospect theory, Quiggin's (1982) rank-dependent utility, and Tversky and Kahneman's (1992) new prospect theory concern the special case of $B(E) = P(E)$, where w now can be nonlinear. The case $B(E) = P(E)$ also includes Gul's (1991) disappointment aversion theory.

We next consider the more general case where no objective probabilities need to be given for all events E. Expected utility is the special case where B is an additive, now "subjective," probability and w is the identity. Choquet expected utility (Schmeidler 1989)

and cumulative prospect theory (Tversky and Kahneman 1992) start from the general weighting function W, from which B obviously results as $w^{-1}(W)$, with w the probability weighting function for risk. The multiple priors model (Gilboa and Schmeidler 1989) results with W(E) the infimum value P(E) over all priors P. Under Machina and Schmeidler's (1992) probabilistic sophistication, B is an additive probability measure.

APPENDIX D. EXPERIMENTAL INSTRUCTIONS

This appendix is not reproduced here, but is made available online.

ACKNOWLEDGMENT. Glenn Harrison and two anonymous referees made helpful comments.

REFERENCES

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

Abdellaoui, M., F. Vossmann, and M. Weber (2005). Choice-based elicitation and decomposition of decision weights for gains and losses under uncertainty. *Management Science* 51, 1384–1399.

Allais, M.(1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine. *Econometrica* 21, 503–546.

Allen, F. (1987). Discovering personal probabilities when utility functions are unknown. *Management Science* 33, 542–544.

Aragones, E., I. Gilboa, A. Postlewaite, and D. Schmeidler (2005). Fact-free learning. *American Economic Review* 95, 1355–1368.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 5, 175–192.

Bleichrodt, H. and J. Luis Pinto (2000). A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management Science* 46, 1485–1496.

Braga, J. and C. Starmer (2005). Preference anomalies, preference elicitation, and the discovered preference hypothesis. *Environmental and Resource Economics* 32, 55–89.

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

Camerer, C.F. and M. Weber (1992). Recent developments in modelling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty* 5, 325–370.

Charness, G. and D.Levin (2005). When optimal choices feel wrong: A laboratory study of Bayesian updating, complexity, and affect. *American Economic Review* 95, 1300–1309.

Chew, S.H. and G. Tan (2005). The market for sweepstakes. *Review of Economic Studies* 72, 1009–1029.

Clemen, R.T. and K.C. Lichtendahl (2005). Debiasing expert overconfidence: A Bayesian calibration model. Fuqua School of Business, Duke University, Durham, NC.

Clemen, R.T. and F. Rolle (2001). In theory … in practice. *Decision Analysis Newsletter* 20, No 1, 3.

de Finetti, B. (1937). La Prévision: Ses Lois Logiques, ses Sources Subjectives. *Annales de l'Institut Henri Poincaré* 7, 1–68.

Echternacht, G.J. (1972). The use of confidence testing in objective tests. *Review of Educational Research* 42, 217–236.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin* 51, 380–417.

Ellsberg, D. (1961). Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics* 75, 643–669.

Ghirardato, P. and M. Marinacci (2001). Risk, ambiguity, and the separation of utility and beliefs. *Mathematics of Operations Research* 26, 864–890.

Gilboa, I. (1987). Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics* 16, 65–88.

Gilboa, I. (2004, Ed.). *Uncertainty in Economic Theory: Essays in Honor of David Schmeidler's 65th Birthday*. Routledge, London.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics* 18, 141–153.

Goeree, J.K., C.A. Holt, and T.R. Palfrey (2002). Quantal response equilibrium and overbidding in private-value auctions. *Journal of Economic Theory* 104, 247–272.

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

Gonzalez, R. and G. Wu (2003). Composition rules in original and cumulative prospect theory. mimeo.

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

Greenspan, A. (2004). Innovations and issues in monetary policy: The last fifteen years. *American Economic Review, Papers and Proceedings* 94, 33–40.

Gul, F. (1991). A theory of disappointment aversion. *Econometrica* 59, 667–686.

Halevy, Y. and V. Feltkamp (2005). A Bayesian approach to uncertainty aversion. *Review of Economic Studies* 72, 449–466.

Halevy, Y. and E. Ozdenoren (2007). Uncertainty and compound lotteries: Calibration. Working paper, University of British Columbia.

Hanson, R. (2002). "Wanna bet? *Nature* 420, November 2002, pp. 354–355.

Harrison, G.W., M.I. Lau, and M.B. Williams (2002). Estimating individual discount rates in Denmark: A field experiment. *American Economic Review* 92, 1606–1617.

Holt, C.A. (1986). Preference reversals and the independence axiom. *American Economic Review* 76, 508–513.

Holt, C.A. (2006). *Webgames and Strategy: Recipes for Interactive Learning*. Addison-Wesley, London, in press.

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

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

Hurwicz, L. (1960). Optimality and informational efficiency in resource allocation. *In* K.J. Arrow, S. Karlin, and P. Suppes (1960, Eds), *Mathematical Methods in the Social Sciences*, 17–46, Stanford University Press, Stanford, CA.

Johnstone, D.J. (2007a). The value of probability forecast from portfolio theory. *Theory and Decision* 63, 153–203.

Johnstone, D.J. (2007b). Economic Darwinism: Who has the best probabilities. *Theory and Decision* 62, 47–96.

Jouini, E. and C. Napp (2007). Consensus consumer and intertemporal asset pricing with heterogeneous beliefs. *Review of Economic Studies* 74, 1149–1174.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica* 47, 263–291.

Karni, E. and Z. Safra (1987). Preference reversal and the observability of preferences by experimental methods. *Econometrica* 55, 675–685.

Keren, G.B. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica* 77, 217–273.

Keynes, J.M. (1921). *A Treatise on Probability.* McMillan, London.

Klibanoff, P., M. Marinacci, and S. Mukerji (2005). A smooth model of decision making under ambiguity. *Econometrica* 73, 1849–1892.

Knight, F.H. (1921). *Risk, Uncertainty, and Profit.* Houghton Mifflin, New York.

Lee, J. (2008). The effect of the background risk in a simple chance improving decision model. *The Journal of Risk and Uncertainty* 36, 19–41.

Li, W. (2007). Changing one's mind when the facts change: Incentives of Experts and the design of reporting protocols. *Review of Economic Studies* 74, 1175–1194.

Luce, R.D. (1959). *Individual Choice Behavior.* Wiley, New York.

Luce, R.D. (1991). Rank- and-sign dependent linear utility models for binary gambles. *Journal of Economic Theory* 53, 75–100.

Luce, R.D. (2000). *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches.* Lawrence Erlbaum Publishers, London.

Maccheroni, F., M. Marinacci, and A Rustichini (2006). Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica* 74, 1447–1498.

Machina, M.J. (2004). Almost-objective uncertainty. *Economic Theory* 24, 1–54.

Machina, M.J. and D. Schmeidler (1992). A more robust definition of subjective probability. *Econometrica* 60, 745–780.

McFadden, D.L. (1974). Conditional logit analysis of qualitative choice behavior. *In* Zarembka (Ed.), *Frontiers of Econometrics*, 105–142, Academic Press, New York.

McFadden, D.L. (1976). Quantal choice analysis: A survey. *Annals of Economic and Social Measurement* 5, 363–390.

Manski, C.F. (2004). Measuring expectations. *Econometrica* 72, 1329–1376.

McClelland, A. and F. Bolger (1994). The calibration of subjective probabilities: Theories and models 1980–1994. *In* G. Wright and P. Ayton (eds.), *Subjective Probability*, 453–481, Wiley, New York.

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

Miyamoto, J.M. (1988). Generic utility theory: Measurement foundations and applications in multiattribute utility theory. *Journal of Mathematical Psychology* 32, 357–404.

Mosteller, F. and P. Nogee (1951). An experimental measurement of utility. *Journal of Political Economy* 59, 371–404.

Murphy, A.H. and R.L. Winkler (1974). Subjective probability forecasting experiments in meteorology: Some preliminary results. *Bulletin of the American Meteorological Society* 55, 1206–1216.

Myagkov, M.G. and C.R. Plott (1997). Exchange economies and loss exposure: Experiments exploring prospect theory and competitive equilibria in market environments. *American Economic Review* 87, 801–828.

Nyarko, Y. and A. Schotter (2002). An experimental study of belief learning using elicited beliefs. *Econometrica* 70, 971–1005.

Olszewski, W. (2007). Preferences over sets of lotteries. *Review of Economic Studies* 74, 567–595.

Nau, R.F. (2006). Uncertainty aversion with second-order utilities and probabilities. *Management Science* 52, 136–145.

Palfrey, T.R. and S.W. Wang (2007). On eliciting beliefs in strategic games. division of the humanities and social sciences, CalTech, Pasadena, CA 91125.

Palmer, T.N. and R. Hagedorn (2006, Eds). *Predictability of Weather and Climate.* Cambridge University Press, Cambridge.

Pfanzagl, J. (1959). A general theory of measurement—applications to utility. *Naval Research Logistics Quarterly* 6, 283–294.

Prelec, D. (1998). The probability weighting function. *Econometrica* 66, 497–527.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science* 306, October 2004, 462–466.

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

Raiffa, H. (1968). *Decision Analysis*. Addison-Wesley, London.

Sandroni, A., R. Smorodinsky, and R.V. Vohra (2003). Calibration with many checking rules. *Mathematics of Operations Research* 28, 141–153.

Savage, L.J. (1954). *The Foundations of Statistics.* Wiley, New York. (2nd edition 1972, Dover Publications, New York.)

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

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

Segal, U. and A. Spivak (1990). First-order versus second-order risk-aversion. *Journal of Economic Theory* 51, 111–125.

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

Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton University Press, NJ.

Shiller, R.J., F. Kon-Ya, and Y. Tsutsui (1996). Why did the Nikkei crash? Expanding the scope of expectations data collection. *The Review of Economics and Statistics* 78, 156–164.

Spiegelhalter, D.J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* 5, 421–433.

Staël von Holstein, C.A.S. (1972). Probabilistic Forecasting: An experiment related to the stock market, *Organizational Behaviour and Human Performance* 8, 139–158.

Starmer, C. and R. Sugden (1991). Does the random-lottery incentive system elicit true preferences? An experimental investigation. *American Economic Review* 81, 971–978.

Tetlock, P.E. (2005). *Expert Political Judgment*. Princeton University Press, NJ.

Thaler, R.H. and E.J. Johnson (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management Science* 36, 643–660.

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

Tversky, A. and D.J. Koehler (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review* 101, 547–567.

van de Kuilen, G., P.P. Wakker, and L. Zou (2008). The midweight method to measure attitudes towards risk and ambiguity. Econometric Institute, Erasmus University, Rotterdam, the Netherlands.

von Neumann, J. and O. Morgenstern (1944, 1947, 1953). *Theory of Games and Economic Behavior.* Princeton University Press, Princeton NJ.

Wakker, P.P. (2004). On the composition of risk preference and belief. *Psychological Review* 111, 236–241.

Wakker, P.P. (2009). *Prospect Theory for Risk and Ambiguity*. Cambridge University Press, forthcoming.

Wakker, P.P. and Daniel Deneffe (1996). Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science* 42, 1131–1150.

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

Wolfers, J. and E. Zitzewitz (2004). Prediction markets. *Journal of Economic Perspective* 18, 107–126.

Wright, W.F. (1988). Empirical comparison of subjective probability elicitation methods. *Contemporary Accounting* 5, 47–57.

Yates, J.F. (1990). *Judgment and Decision Making.* Prentice Hall, London.