Correcting Proper Scoring Rules for Risk Attitudes[1]

Theo Offerman[a], Joep Sonnemans[a], Gijs van de Kuilen[a], & Peter P. Wakker[b]

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

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

May, 2006

ABSTRACT. Proper scoring rules, introduced in the 1950s, efficiently elicit subjective beliefs about likelihoods, but do so only under the assumption of expected value maximization. The latter assumption can be violated because of nonlinear utility (Bernoulli), nonexpected utility (Allais), and ambiguity attitudes for unknown probabilities (Keynes, Knight, Ellsberg). These violations of expected value have been incorporated in modern decision theories. This paper shows how proper scoring rules can be generalized to those modern theories, and can become valid under risk aversion and other deviations from expected value. An experiment demonstrates the feasibility of our extension, yielding plausible empirical results. Violations of additivity of subjective probabilities are reduced by our extension, although they do not disappear entirely, which suggests genuine nonaddivity in subjective beliefs. The quality of reported probabilities is better under repeated small incentives than under single large incentives.

Key words: belief measurement, proper scoring rules, ambiguity, Knightian uncertainty, subjective probability, nonexpected utility

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

1. Introduction

In many situations, no probabilities are known of uncertain events that are relevant to our decisions, and subjective assessments of the likelihoods of such events have to be made. Proper scoring rules provide an efficient tool for eliciting such subjective assessments from choices. They use cleverly constructed optimization problems where the observation of one single choice suffices to determine the exact quantitative degree of belief of an agent. This procedure is more efficient than the observation of binary choices, as is most common in decision theory, because binary choices only give inequalities and approximations of beliefs.

The measurement of subjective beliefs is important in many domains (Gilboa & Schmeidler 1999; Manski 2004), and proper scoring rules have been widely used accordingly, in accounting (Wright 1988), Bayesian statistics (Savage 1971), business (Staël von Holstein 1972), education (Echternacht 1972), medicine (Spiegelhalter 1986), psychology (Liberman & Tversky 1993; McClelland & Bolger 1994), and other fields (Hanson 2002, Prelec 2004). Proper scoring rules are especially useful for giving experts incentives to exactly specify their degrees of belief. They are commonly used, for instance, to improve the calibration of weather forecasters (Palmer & Hagedorn 2006). They have recently become popular in experimental economics and game theory. Advocates of the frequentist interpretation of probability can become more interested in subjective probabilities when exposed to proper scoring rules. The quadratic scoring rule is the most popular proper scoring rule today (McKelvey & Page 1990; Nyarko & Schotter 2002), and is the topic of this paper.

Proper scoring rules were introduced independently by Brier (1950), Good (1952, p. 112), and de Finetti (1962), and are based on the assumption of expected value maximization, i.e. risk neutrality. All applications up to today that we are aware of have maintained this assumption. Empirically, however, deviations from expected value maximization are common. First, Bernoulli (1738) pointed out that risk aversion prevails over expected value, so that, under expected utility, utility has to be concave rather than linear. Second, Allais (1953) demonstrated, for events with known probabilities, that people can be risk averse in ways that expected utility cannot accommodate, so that more general decision theories are called for with other factors in addition to utility curvature (Kahneman & Tversky 1979, Machina 1982, Quiggin 1982, Tversky & Kahneman 1992). Third, Keynes (1921), Knight (1921), and Ellsberg (1961) demonstrated the importance of ambiguity for events with unknown

probabilities ("Knightian uncertainty"), where phenomena occur that are fundamentally different than those for known probabilities, adding to the descriptive failure of expected value and showing that there are further factors generating deviations from expected value. Gilboa (1987), Gilboa & Schmeidler (1989), Schmeidler (1989), and Tversky & Kahneman (1992) developed decision theories that incorporate ambiguity.

This paper updates 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, on the one hand, we extend the applicability and validity of proper scoring rules to the empirical phenomena put forward by modern decision theorists. On the other hand, we extend the applicability of those modern decision theories by introducing the efficient measurement technique of proper scoring rules into those theories. We give quantitative assessments of the distortions of classical proper scoring rules caused by the empirical deviations from expected value, and show how such distortions can be corrected for by means of our generalized method.

Our correction technique can be interpreted as a new method of calibration (Keren 1991, Yates 1990) that does not need many repeated observations, unlike traditional calibration techniques (Clemen & Lichtendahl 2002, Leher 2001). An efficient aspect of our method is that we need not elicit entire risk attitudes of agents so as to correct for them. For instance, for nonexpected utility we need not go through an entire measurement of the utility and probability weighting functions to apply our correction. Instead, we can immediately infer the correction from a limited set of readily observable data. We demonstrate the feasibility of our method in an experiment where we measure the subjective beliefs of participants about the future performance of stocks after provision of information about past performance. The empirical findings will confirm the usefulness of our method.

The first part of this paper (Sections 2-5) is theoretical and the second part (Sections 6 and further) is experimental. Section 2 provides the definition of proper scoring rules, and Section 3 provides the classical analysis of proper scoring rules, which assumes expected value maximization. A decision-theoretic analysis of the three main deviations from expected value maximization, and their effects on proper scoring rules, is in Section 4. Section 5 presents risk-corrections, serving to correct for the biases described in Section 4. Section 6 contains a simple example to illustrate how the theory developed in the first five sections can be applied empirically. It shows in particular that many decision-theoretic details, presented in the first part to justify our correction procedures, need not be studied when applying our method empirically. Readers interested in applying our method

empirically can skip most of Sections 3-5, reading only Section 3 up to Theorem 3.1 and Corollary 5.1. Methodological details of our experiment are described in Section 7. Section 8 presents results regarding the biases that we correct for, and Section 9 some implications of the corrections of such biases. Discussions and conclusions are in Sections 10 and 11. Appendix A presents proofs and technical results, Appendix B surveys the implications of modern decision theories for our measurements, and Appendix C presents details of the experimental instructions.

2. Proper Scoring Rules; Definitions

Let E denote an event of which an agent is uncertain about whether or not it obtains, such as snow in Amsterdam in March 1932, whether a stock's value will decrease during the next half year, whether a ball randomly drawn from 20 numbered balls will have a number below 5, whether the 100^{th} digit of π is 3, etc. The degree of uncertainty of an agent about E will obviously depend on the information that the agent has about E. Some agents may even know with certainty about some of the events. Most events will, however, be uncertain. For most uncertain events, no objective probabilities of occurrence are known, and our decisions have to be based on subjective assessments, consciously or not, of their likelihood.

Prospects designate event-contingent payments. We use the general notation (E:x, 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. The unit of payment for outcomes is one dollar. *Risk* concerns the case of known probabilities. Here, for a prospect (E:x, y), the probability p of event E is known, and we can identify this prospect with a probability distribution (p:x, y) over money, yielding x with probability p and y with probability 1−p.

Several methods have been used in the literature to measure the subjective degree of belief of an agent in event E.

- In decision theory, binary choices have been used for this purpose. If the agent prefers the prospect (E:1, 0) to receiving 0.10 for sure and also to receiving 0.20 for sure, but rather receives 0.30 for sure than the prospect, then under some assumptions (expected value; see later) we may conclude that the subjective probability of E is between 0.20 and 0.30. In this way, many choices must be observed to infer levels of belief, and only approximations are obtained.

- We may seek to observe an indifference between (E:1, 0) and a sure gain r, in which case we can conclude (again assuming expected value) that the degree of belief in E is exactly r. It is, however, hard to obtain measurements of indifferences. A Becker-DeGroot-Marschak mechanism can be used. Some drawbacks of this method are that it is fairly complex (Albers, Pope, Selten, & Vogt 2000, p. 116; Braga & Starmer 2005), needs further assumptions (Becker, de Groot, & Marschak 1964, p. 226; Guala 2000; Karni & Safra 1987), and is prone to irrational auction strategies (Plott & Zeiler 2005 p. 537).
- de Finetti (1931, 1937) devised a clever book-making technique with agents having to announce subjective probabilities, others taking them up on the corresponding betting odds, and the agents vulnerable to money pumps as soon as they violate the Bayesian laws of probability. This technique, however, concerns hypothetical decision situations and cannot be implemented in practice.
- Introspective questions can be used, simply asking what the degree of belief r is. It is, however, not clear why an agent would answer these questions truthfully and, without incentives, it may even be unclear to the agent (e.g. when the agent is a frequentist) what this number r is supposed to signify.

Proper scoring rules provide an efficient and operational manner for measuring subjective beliefs that deliver what the above methods seek to do, while avoiding the problems mentioned. They do not consider binary choice as in most decision theories but instead a multiple-choice optimization problem. They ask for the value r above without need to observe indifferences and with proper incentives added, and have been widely implemented unlike the book-making devise.

Under the *quadratic scoring rule* (*QSR*), the most commonly used proper scoring rule and the rule considered in this paper, a *qsr-prospect*

$$(E: 1-(1-r)^2,\ 1-r^2), \tag{2.1}$$

is offered to the agent, where $0 \le r \le 1$ is a number that the agent can choose freely. The number chosen is a function of E, sometimes denoted r_E, and is called the (*uncorrected*) *reported probability* of E. The reasons for this term will be explained later. More general prospects $(E: a-b(1-r)^2,\ a-br^2)$ for any $b>0$ and $a \in \mathbb{R}$ can be considered, but for simplicity we restrict our attention to $a = b = 1$. No negative payments can occur, so that the agent never loses money. It is obvious that if the agent is certain that E will obtain, then he will maximize $1-(1-r)^2$, irrespective of $1-r^2$, and will choose $r=1$. Similarly, $r=0$ is chosen if E

will certainly not obtain. The choice of r = 0.5 gives a riskless prospect, yielding 0.75 with certainty. Increasing r increases the payment under E but decreases it under E^c. Under the event that happens, the QSR pays 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 (r under E, 1–r under E^c).

We, finally, discuss a symmetry between E and E^c, leading to a restriction on the data that can be obtained through QSRs and that will be further discussed in Subsection 4.3. For any prospect, the agent can have the payments under E and E^c interchanged by interchanging r and 1–r. We display this point for future reference.

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 reported probability of E identical to the prospect resulting from 1–r as reported probability of E^c. □

Because of Observation 2.1, we have

$$r_{E^c} = 1 - r_E. \tag{2.2}$$

This symmetry will hold throughout this paper, and will be crucial for the nonexpected utility analysis presented later.

3. Proper Scoring Rules and Subjective Expected Value

When proper scoring rules were introduced in the 1950s, they were based on the assumption of subjective expected value maximization. This assumption is still made in applications of proper scoring rules today, and will also be made in this section. It means, first, that the agent assigns a subjective probability p to each event E.[2] Under common

[2] In this paper, the term *subjective probability* is used only for probability judgments that are Bayesian in the sense of satisfying 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. Such concepts, different than

richness assumptions, the subjective probabilities must agree with objective probabilities whenever the latter are given to the agent[3], and such agreement is assumed throughout this paper. Second, the agent maximizes expected value with respect to probabilities, irrespective of whether these are objective or subjective. These assumptions together constitute the *subjective expected value* model.

For QSRs and an event E with probability p, subjective expected value implies that the agent maximizes

$$p \times (1-(1-r)^2) + (1-p) \times (1-r^2) = 1 - p(1-r)^2 - (1-p)r^2. \quad (3.1)$$

If event E has probability p, then we also write R(p) for r_E throughout this paper. According to Eq. 3.1, and all other models considered in this paper, all events E with the same probability p have the same value r_E, so that R(p) is well-defined. We have the following corollary of Eq. 2.2.

$$R(1-p) = 1 - R(p). \quad (3.2)$$

The following theorem demonstrates that the QSR is incentive compatible. The theorem immediately follows from the first-order optimality condition $2p(1-r) - 2r(1-p) = 0$ in Eq. 3.1. Second-order optimality conditions are verified throughout this paper and will not be mentioned in what follows.

THEOREM 3.1. Under subjective expected value maximization, the optimal choice r_E is equal to the (subjective or objective) probability p of event E, i.e. R(p) = p. □

It is in the agent's best interest to truly report his subjective probability of E. This explains the term "reported probability." Reported probabilities satisfy the Bayesian additivity condition for probabilities. We call the number r_E the (*uncorrected*) *reported probability*.

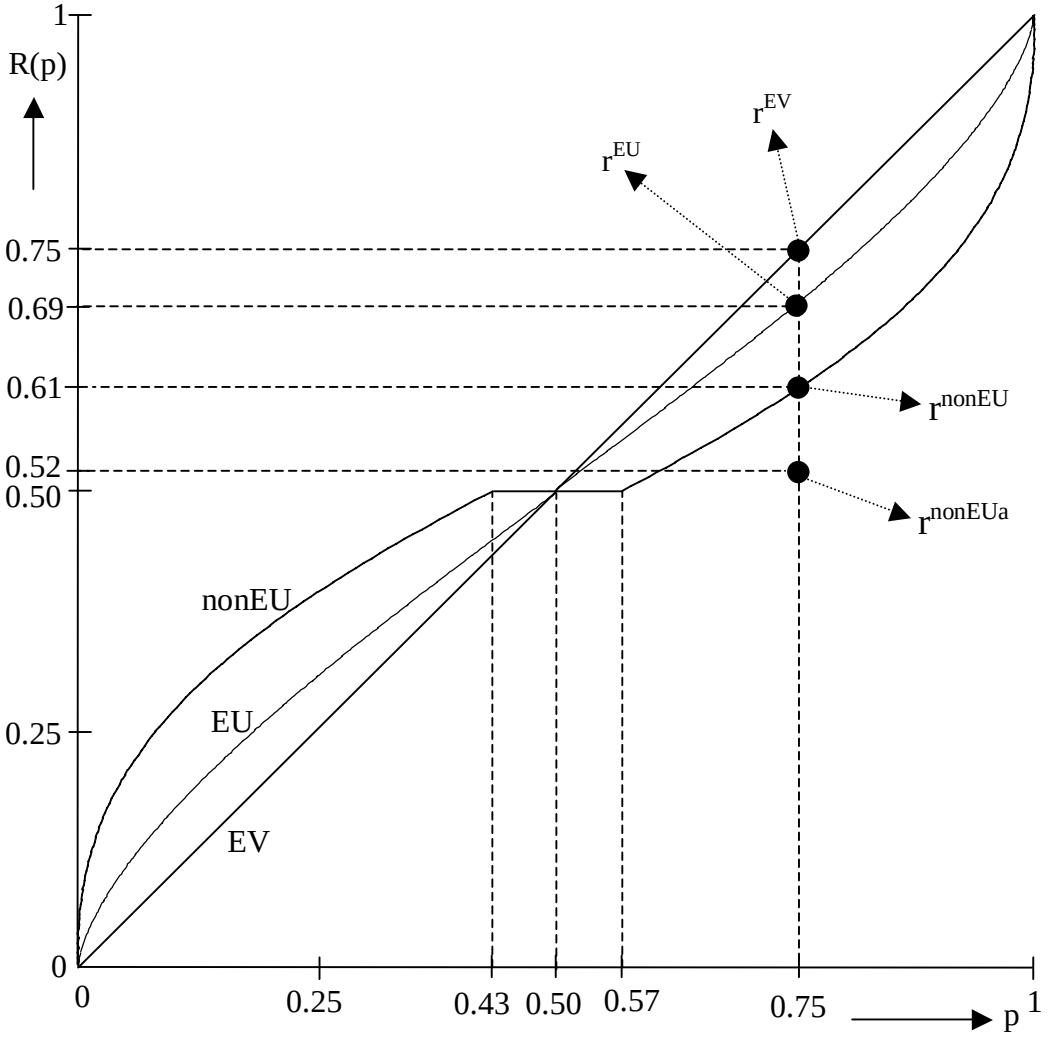
---

probabilities, will be analyzed in later sections, and we will use the term (probability) weights or beliefs, depending on the way of generalization, to designate them.

[3] This follows first for equally-probable n-fold partitions of the universal event, where because of symmetry all events must have both objective and subjective probabilities equal to 1/n. Then it follows for all events with rational probabilities because they are unions of the former events. Finally, it follows for all remaining events by proper continuity or monotonicity conditions. There have been several misunderstandings about this point, especially in the psychological literature (Edwards 1954, p. 396; Schoemaker 1982, Table 1).

Figure 3.1 depicts R(p) as a function of the probability p which, under expected value as considered here, is simply the diagonal r = p, indicated through the letters EV. The other curves and points in the figure will be explained later. Throughout the first part of this paper, we use variations of the following theoretical example.

FIGURE 3.1. 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 Example 3.1.3.
r^{nonEUa}: nonexpected utility for unknown probabilities ("Ambiguity").

EXAMPLE 3.2. An urn K ("known" distribution) contains 25 Crimson, 25 Green, 25 Silver, and 25 Yellow balls. One ball will be drawn at random. C designates the event of a crimson ball drawn, and G, S, and Y are similar. E is the event that the color is not crimson, i.e. it is the event $C^c = \{G,S,Y\}$. Under expected value maximization, $r_E = R(0.75) = 0.75$ is optimal

in Eq. 2.1, yielding prospect (E:0.9375, 0.4375) with expected value 0.8125. The point r_E is depicted as r^{EV} in Figure 3.1. Theorem 3.1 implies that $r_G = r_S = r_Y = 0.25$. We have $r_G + r_S + r_Y = r_E$, and the reported probabilities satisfy additivity. □

4. Three Commonly Found Deviations from Subjective Expected Value and Their Implications for Quadratic Proper Scoring Rules

This section describes three factors that generate deviations from expected value maximization and, hence, can distort the classical analyses of proper scoring rules. The effects of each factor in this section are illustrated in Figure 3.1, explained later, and their quantitative size will be illustrated through extensions of Example 3.2. Subsection 4.1 considers the first factor generating deviations, being nonlinear utility under expected utility. This section extends an earlier study of this factor by Winkler & Murphy (1970). Bliss & Panigirtzoglou (2004) also corrected estimations of probability distributions for utility curvature, but did not do it in the context of proper scoring rules. Subsection 4.2 considers the second factor, i.e. violations of expected utility for known probabilities. Subsection 4.3 considers the third factor, i.e. ambiguity because of unknown probabilities.

4.1. The First Deviation: Utility Curvature

Bernoulli (1738) put forward the first deviation from expected value. Because of the so-called St. Petersburg paradox, Bernoulli proposed that people maximize expected utility with respect to a *utility function* U, which we assume continuously differentiable with positive derivative everywhere, implying strict increasingness. We assume throughout that U(0) = 0. Eq. 3.1 is now generalized to

$$pU(1-(1-r)^2) + (1-p)U(1-r^2)\,. \qquad (4.1.1)$$

The first-order optimality condition for r, and a rearrangement of terms (as in the proof of Theorem 4.3.2), implies the following result. For $r \neq 0.5$, the theorem also follows as a corollary of Theorem 4.3.2 and Eq. 3.2.

THEOREM 4.1.1. Under expected utility with p the (subjective or objective) probability of event E, the optimal choice r = R(p) satisfies:

$$r = \frac{p}{p + (1-p)\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}} \quad . \tag{4.1.2}$$

□

A utility correction is imposed, based on the marginal-utility ratio at the two prizes of the qsr-prospect. It implies that r deviates from the objective probability p if the marginal utilities at the two prizes are different. For concave utility and r > 0.5, so that E is judged more probable than E^c and receives the highest payment, the marginal-utility ratio will exceed 1 and r will be lower, and closer to 0.5, than p. For r < 0.5, r will also be closer to 0.5 than p, because 1–r, the reported probability of E^c, is so too. It follows that risk aversion moves people in the direction of the riskless prospect of r = 0.5 (Kadane & Winkler 1988, p. 359), a phenomenon confirmed empirically by Huck & Weiszäcker (2002) and Winkler (1967).

Figure 3.1 depicts an example of the function r under expected utility, indicated by the letters EU, and is similar to Figure 3 of Winkler & Murphy (1970). The decision-based distortion in the direction of 0.5 is opposite to the overconfidence (probability judgments too far from 0.5) found mostly in direct judgments of probability without real incentives (Fischer 1982; Fischhoff, Slovic & Lichtenstein 1977), and found among experts seeking to distinguish themselves (Keren 1991, p. 224 and 252; the "expert bias", Clemen & Rolle 2001).

EXAMPLE 4.1.2. Consider Example 3.2, but assume expected utility with $U(x) = x^{0.5}$. Substitution of Eq. 4.1.2 (or Theorem 4.3.2 below) shows that $r_E = R(0.75) = 0.69$ is optimal, depicted as r^{EU} in Figure 3.1, and yielding prospect (E:0.91, 0.52) with expected value 0.8094. The extra risk aversion generated by concave U has led to a decrease of r_E by 0.06 relative to Example 3.2, distorting the probability elicited, and generating an expected-value loss of 0.8125 – 0.8094 = 0.0031. This amount can be interpreted as an uncertainty premium, designating the profit margin for an insurance company. The uncertainty premium will be larger for deviations from expected value considered later. 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, because r_G

297 $+ r_S + r_Y = 0.93 > 0.69 = r_E$. This violation in the data reveals that expected value does not
298 hold. □
299
300 The above example illustrates how observations of reported probabilities can be used to
301 directly reveal violations of additivity empirically. We are not aware of such tests in the
302 literature. Under such violations, reported probabilities may not truly reveal beliefs but may
303 be distorted by other factors. We will report empirical tests of additivity in Sections 6 and 9.
304
305 OBSERVATION 4.1.3. Under expected utility with probability measure P, $r_E = 0.5$ implies
306 $P(E) = 0.5$. Conversely, $P(E) = 0.5$ implies $r_E = 0.5$ if risk aversion holds. Under risk
307 seeking, $r_E \neq 0.5$ is possible if $P(E) = 0.5$. □
308
309 Theorem 4.1.1 clarifies the distortions generated by nonlinear utility, but it does not
310 provide an explicit expression of R(p), i.e. r as a function of p, or vice versa. It seems to be
311 impossible, in general, to obtain an explicit expression of R(p). We can, however, obtain an
312 explicit expression of the inverse of R(p), i.e. p in terms of r. For numerical purposes, R(p)
313 can then be obtained as the inverse of that function—this is what we did in our numerical
314 analyses, and how we drew Figure 3.1. The following result follows from algebraic
315 manipulations or, for $r \neq 0.5$, as a corollary of Corollary 4.2.5 hereafter.
316
317 COROLLARY 4.1.4. Under Eq. 4.1.2, the optimal choice r = R(p) satisfies:

318
$$p = \frac{r}{r + (1-r)\frac{U'(1-(1-r)^2)}{U'(1-r^2)}} . \qquad (4.1.3)$$

319 □
320 The result shows that the relation between p and r is nonlinear, so that r will violate
321 additivity, as soon as marginal utility U´ is nonsymmetric about 0.5, which holds for all
322 regular nonlinear utility functions. Hence, additivity of reported probability provides a
323 critical test for linearity of utility under expected utility.
324

4.2. The Second Deviation: Nonexpected Utility for Known Probabilities

In the nonexpected utility analyses that follow, we will often restrict our attention to $r \geq 0.5$. Results for $r < 0.5$ then follow by interchanging E and E^c, and the symmetry of Observation 2.1 and Eq. 2.2.

We say that event A is (*revealed*) *more likely than* event B if, for some positive outcome x, say x = 100, the agent prefers (A:x, 0) to (B:x, 0). In all models considered hereafter, 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 4.2.1. 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 logarithmic proper scoring rules, as well as all other proper scoring rules as they have been applied in the literature so far.

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 be concluded then. 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 $r > 0.5$ combined with E being strictly less likely than E^c are excluded for QSRs.

We now turn to the second deviation from de Finetti's assumption of expected value, put forward by Allais (1953), which deviates from Bernoulli's expected utility, and still pertains to events E with known probability p. With M denoting 10^6, the preferences $M \succ$ (0.8: 5M, 0) and (0.25:M, 0) $\prec$ (0.20:5M, 0) are plausible. They would imply, under expected utility with U(0) = 0, the contradictory inequalities $U(M) > 0.8 \times U(5M)$ and $0.25U(M) < 0.20 \times U(5M)$ (implying $U(M) < 0.8 \times U(5M)$), so that they falsify expected utility. It has since been shown that this paradox does not concern an exceptional phenomenon pertaining only to hypothetical laboratory choices with extreme amounts of money, but that the phenomenon is relevant to real decisions for realistic stakes (Kahneman & Tversky 1979). The Allais paradox and other violations of expected utility have led to several alternative models for decision under risk, the so-called nonexpected utility models.

For the prospects relevant to this paper, QSRs with only two outcomes and no losses, all currently popular nonexpected-utility evaluations of qsr-prospects (Eq. 2.1) are of the following form (see Appendix B). We first present such evaluations for the case of highest payment under event E, i.e. $r \geq 0.5$, which can be combined with $p \geq 0.5$.

For $r \geq 0.5$: $w(p)U(1-(1-r)^2) + (1-w(p))U(1-r^2)$. (4.2.1)

Here w is a continuous strictly increasing function with $w(0) = 0$ and $w(1) = 1$, and is called a *probability weighting function*. Expected utility is the special case of $w(p) = p$. By symmetry, the case $r < 0.5$ corresponds with a reported probability $1-r > 0.5$ for E^c, giving the following representation.

For $r < 0.5$: $w(1-p)U(1-r^2) + (1 - w(1-p))U(1-(1-r)^2)$. (4.2.2)

The different weighting of an event when it has the highest or lowest outcome is called rank-dependence. It suffices, by Eqs. 2.2 and 3.2, to analyze the case of $r \geq 0.5$ for all events.

Both in Eq. 4.2.1 and in Eq. 4.2.2, w is applied only to probabilities $p \geq 0.5$, and needs to be assessed only on this domain in what follows. This restriction is caused by Observation 4.2.1. We display the implication.

OBSERVATION 4.2.2. 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. □

Hence, for the risk-correction introduced later, we need to estimate w only on [0.5,1]. An advantage of this point is that the empirical findings about w are uncontroversial on this domain, the general finding being that w underweights probabilities there.[4] Under nonexpected utility, not only a utility correction must be imposed, but also a probability weighting correction w must be applied to p, leading to the following result.

THEOREM 4.2.3. Under nonexpected utility with p the probability of event E, the optimal choice $r = R(p)$ satisfies:

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

For $r > 0.5$: $r = \dfrac{w(p)}{w(p) + (1-w(p))\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}$. (4.2.3)

□

The above result, again, follows from the first-order optimality condition, and also follows as a corollary of Theorem 4.3.2 below. As an aside, the theorem shows that QSRs provide an efficient manner for measuring probability weighting on (0.5, 1] if utility is linear, because then simply $r = R(p) = w(p)$. An extension to [0, 0.5] can be obtained by a modification of QSRs, discussed further in the next subsection (Eqs. 4.3.3 and 4.3.4).

EXAMPLE 4.2.4. Consider Example 4.1.2, but assume nonexpected utility with $U(x) = x^{0.5}$ and

$$w(p) = \left(exp(-(-ln(p))^{\alpha})\right) \quad (4.2.4)$$

(Prelec 1998), with parameter $\alpha = 0.65$. This function agrees with common empirical findings (Tversky & Kahneman 1992, Abdellaoui 2000, Bleichrodt & Pinto 2000, Gonzalez & Wu 1999). From Theorem 4.2.3 it follows that $r_E = R(0.75) = 0.61$ is now optimal, depicted as r^{nonEU} in Figure 3.1. It yields prospect (E:0.85, 0.63) with expected value 0.7920. The extra risk aversion relative to Example 4.1.2 generated by w for this event E has led to an extra distortion of r_E by 0.08. The extra expected-value loss (uncertainty premium) relative to Example 4.1.2 is 0.8094 – 0.7920 = 0.0174. By Eq. 4.2.1, $r_C = 0.39$, 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 effects of probability weighting are strongest near $p = 0.75$ and, indeed, relative to Example 4.1.2, nonexpected utility has generated a large extra distortion in the above example. Figure 3.1 illustrates the effects through the curve indicated by nonEU. Note that the curve is flat around $p = 0.5$, 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). Such a degree of risk aversion is not possible under expected utility, where $r = 0.5$ can happen only for $p = 0.5$ (Observation 4.1.3). 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 neighborhood of p = 0.5. An explicit expression of p in terms of r, i.e. of $R^{-1}(p)$, follows next for r > 0.5, assuming that we can invert w.

COROLLARY 4.2.5. Under Eq. 4.2.1, the optimal choice r = R(p) satisfies:

$$\text{If } r > 0.5, \text{ then } p = R^{-1}(r) = w^{-1}\left(\frac{r}{r + (1-r)\frac{U'(1-(1-r)^2)}{U'(1-r^2)}}\right). \qquad (4.2.5)$$

□

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. 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 4.1.3).

Up to this point, we considered deviations from expected value and Bayesianism at the level of decision attitude, and beliefs themselves were not yet affected. This will be different in the next subsection.

4.3. The Third Deviation: Nonadditive Beliefs and Ambiguity for Unknown Probabilities

This section considers events for which no probabilities are known. It is commonly assumed in applications of proper scoring rules that the agent then chooses (Bayesian) subjective probabilities p = P(E) for such events E, satisfying the laws of probability, and evaluates prospects the same way for subjective probabilities as if these probabilities were objective. Such an approach to unknown probabilities, staying as close as possible to risk, is called *probabilistic sophistication* (Machina & Schmeidler 1992). In traditional applications of proper scoring rules it is further assumed that the agent satisfies expected utility for known probabilities, but probabilistic sophistication is more general and allows deviations like those in the preceding subsection for known probabilities. Probabilistic sophistication can be interpreted as a last attempt to at least maintain Bayesianism at the level of beliefs. As we will see next, however, it fails descriptively. Empirical findings, initiated by Ellsberg (1961),

have demonstrated that probabilistic sophistication is commonly violated empirically. The following example gives details. For another kind of violation see Marinacci (2002).

EXAMPLE 4.3.1 [violation of probabilistic sophistication]. Consider Example 4.2.4, but now there is an additional urn A ("ambiguous"). Like urn K, A contains 100 balls colored Crimson, Green, Silver, or Yellow, but now the proportions of balls with these colors are unknown. C_a designates the event of a crimson ball drawn from A, and G_a, S_a, and Y_a are similar. E_a is the event $C_a^c = \{G_a, S_a, Y_a\}$. If probabilities are assigned to drawings from the urn A (as assumed by probabilistic sophistication) then, in view of symmetry, we must have $P(C_a) = P(G_a) = P(S_a) = P(Y_a)$, so that these probabilities must be 0.25. Then $P(E_a)$ must be 0.75, as was P(E) in Example 4.2.4. Under probabilistic sophistication combined with nonexpected utility as in Example 4.2.4, r_{E_a} must be the same as r_E in Example 4.2.4 for the known urn, i.e. $r_{E_a} = 0.61$. It implies that people must be indifferent between (E:x, y) and (E_a:x, y) for all x and y. The latter condition is typically violated empirically. People usually have a strict preference for known probabilities, i.e. (E:x, y) $\succ$ (E_a:x, y).[5] Consequently, it is impossible to model beliefs about uncertain events E_a through probabilities, and probabilistic sophistication must fail. This observation also suggests that r_{E_a} may differ from r_E. □

The deviations from expected value illustrated by the above example cannot be explained by utility curvature or probability weighting, and must be generated by other factors. Those other, new, factors refer to properties of beliefs and decision attitudes that are typical of unknown probabilities, and force us to give up on the additive measure P(E) in our model. Besides decisions, also beliefs may deviate from the Bayesian principles.

The important difference between known and unknown probabilities was first emphasized by Keynes (1921) and Knight (1921). Keynes discussed the example of urns with unknown compositions. Demonstrations as in the above example, proving that it is impossible to account for observed behavior in terms of probabilities, were first given by Ellsberg (1961).

Studies of direct judgments have supported the thesis that subjective beliefs may deviate from Bayesian probabilities (Dempster 1968; McClelland & Bolger 1994; Shafer 1976; Tversky & Koehler 1994). Instead of a probability p of E, we have to substitute a general

[5] This holds also if people can choose the three colors to gamble on in the ambiguous urn, so that there is no reason to suspect unfavorable compositions.

subjective function B(E) in Eq. 4.2.1, with Eq. 4.2.2 adapted similarly, and with nonadditivity of B adding to the deviations from Bayesianism. B may be interpreted as a belief index, as it was in Schmeidler (1989) who initiated the use of nonadditive measures for unknown probabilities. There is no consensus in decision theory today about whether B can also comprise other components of decision attitude beyond beliefs.

In Example 4.3.1 we may have $B(C_a) = B(G_a) = B(S_a) = B(Y_a) < 0.25$, $B(E_a) < 0.75$, and $B(E_a) \neq B(G_a) + B(S_a) + B(Y_a)$. Such phenomena lead to the following general evaluation of the qsr-prospects of Eq. 2.1, for general events E.

$$\text{For } r \geq 0.5\text{: } w(B(E))U(1-(1-r)^2) + (1-w(B(E)))U(1-r^2). \tag{4.3.1}$$

The evaluation for $r < 0.5$ can again be obtained from Eq. 2.2. We give it for completeness.

$$\text{For } r < 0.5\text{: } (1-w(B(E^c)))U(1-(1-r)^2) + w(B(E^c))U(1-r^2). \tag{4.3.2}$$

In general, B assigns value 0 to the vacuous event $\varnothing$, 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 obviously also hold for the composition $w(B(\cdot))$. This composition is called the *weighting function*, and is denoted W. In the literature, the weighting function W is usually taken as point of departure.[6] Given W and strict increasingness of w, we can define $B = w^{-1}(W)$ so as to obtain consistency of notation.

The equality $B(E) + B(E^c) = 1$ (*binary additivity*) may very well be violated. Then it can be debated whether $B(E)$ or $1 - B(E^c)$, or some other index, is to be taken as index of belief, and whether other decision components beyond beliefs are comprised in some or all of these indexes. Such interpretations have not yet been settled, and further studies are called for. Whereas the interpretation of B as index of belief is open to debate, depending on further developments in decision theory, it is not open to debate that the behavioral component of risk attitude should be filtered out before an interpretation of belief can be considered. Filtering out this behavioral component, as a necessary preparation for further investigations of beliefs, is the contribution of this paper.

---

[6]Schmeidler (1989) and most other current studies of uncertainty assume, for simplicity, that w is the identity. Then B and W coincide.

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 above equations, reflecting again the restriction of the QSR of Observation 4.2.1. 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

$$(E: 2-(1-r)^2,\ 1-r^2), \tag{4.3.3}$$

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 Section 2 are not affected by this modification, and the results of Section 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 restriction of Observation 4.2.1. We then can observe w of the preceding subsection, and W(E) and B(E) over their entire domain. Similarly, with prospects

$$(E: 1-(1-r)^2,\ 2-r^2), \tag{4.3.4}$$

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 confine our attention to the QSRs of Eq. 2.1 as they are classically applied throughout the literature, so as to reveal their biases according to the current state of the art of decision theory, suggesting remedies whenever possible, and signaling the problems that remain. We leave further investigations of the, we think promising, modifications of QSRs in the above equations to future studies.

The restrictions of the classical QSRs will also hold for the experiment reported later in this paper. There an application of the QSR to a small interval I is to be interpreted formally as the measurement of $1 - B(I^c)$. The restrictions also explain why the theorems below concern only the case of $r > 0.5$ (with $r = 0.5$ as a boundary solution).

The following theorem, our main theorem, specifies the first-order optimality condition for interior solutions of r for general decision making, incorporating all deviations described above.

THEOREM 4.3.2. Under Eq. 4.3.1, the optimal choice r satisfies:

$$\text{If } r > 0.5, \text{ then } r = r_E = \frac{w(B(E))}{w(B(E)) + (1-w(B(E)))\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}\ . \tag{4.3.5}$$

□

We cannot draw graphs as in Figure 3.1 for unknown probabilities, because the x-axis now concerns events and not numbers. The W values of ambiguous events will be relatively low for an agent with a general aversion to ambiguity, so that the reported probabilities r in Eq. 4.3.5 will be relatively small, i.e. close to 0.5. We give a numerical example.

EXAMPLE 4.3.3. Consider Example 4.3.1. Commonly found preferences (E:100, 0) $\succ$ (E_a:100, 0) imply that $w(B(E_a)) < w(B(E)) = w(0.75)$. Hence, by Theorem 4.3.2, r_{E_a} will be smaller than r_E. Given the strong aversion to unknown probabilities that is often found empirically (Becker & Brownson 1964; Camerer & Weber 1992), we will assume that $r_{E_a} = 0.52$. It is depicted as r^{nonEUa} in Figure 3.1, and yields prospect (E_a:0.77, 0.73) with expected value 0.7596. The extra preference for certainty relative to Example 4.2.4 generated by unknown probabilities for this event E_a has led to an extra distortion of r_{E_a} by $0.61 - 0.52 = 0.09$. The extra expected-value loss relative to Example 4.2.4 is $0.7920 - 0.7596 = 0.0324$. This amount can be interpreted as the ambiguity-premium component of the total uncertainty premium. By Eq. 4.2.1, $r_C = 0.48$, and by symmetry $r_G = r_S = r_Y = 0.48$ too. The reported probabilities violate additivity to an extreme degree, because $r_G + r_S + r_Y = 1.44 > 0.52 = r_{E_a}$. The behavior of the agent is close to a categorical fifty-fifty evaluation, where all nontrivial uncertainties are weighted the same without discrimination.

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, however, equal $B\{G_a,S_a,Y_a\} = B(E_a)$ under additivity which is 0.62, leading to a contradiction. Hence, additivity must be violated.

Of the total deviation of $r_{E_a} = 0.52$ from 0.75, being 0.23, a part of $0.06 + 0.08 = 0.14$ is the result of deviations from risk neutrality that distorted the measurement of $B(E_a)$, and 0.09 is the result of nonadditivity (ambiguity) of belief B. □

Theorem 4.3.2 is valid for virtually all models of decision under uncertainty and ambiguity currently known in the literature, because Eqs. 4.3.1 and 4.3.2 capture all these models (see Appendix B). Some qualitative observations are as follows. If U is linear, then r = w(B(E)) follows for all $w(B(E)) > 0.5$, providing a very tractable manner of measuring the

nonadditive decision-theory measure $W = w \circ B$. A corner solution $r = 0.5$ results for all $w(B(E)) \leq 0.5$ with also $w(B(E^c)) \leq 0.5$, so that the classical QSR has no discriminatory power for such events. For events with known probabilities, such corner solutions correspond with the flat part of the nonEU curve in Figure 3.1. For ambiguous events, ambiguity aversion will enhance the existence of such corner solutions.

Expected utility is the special case where $W(B(E)) = P(E)$ for a probability measure P, so that Eqs. 4.3.1 and 4.3.2 are the same and each applies to all r. Theorem 4.1.1 demonstrated that Eq. 4.3.5 then also holds for $r \leq 0.5$.

In applications of proper scoring rules, we usually first observe $r = r_E$ and then want to derive B(E) from r. The following corollary gives an explicit expression. It illustrates once more how deviations from expected utility (w) and nonlinear utility (the marginal-utility ratio) distort the classical proper-scoring-rule assumption of $B(E) = r$.

COROLLARY 4.3.4. Under Eq. 4.3.1, the optimal choice $r = r_E$ satisfies:

$$\text{If } r > 0.5, \text{ then } B(E) = w^{-1}\left(\frac{r}{r + (1-r)\frac{U'(1-(1-r)^2)}{U'(1-r^2)}}\right). \qquad (4.3.6)$$

□

5. Measuring Beliefs through Risk Corrections

One way to measure B(E) is by eliciting W(E) and the function w from choices under uncertainty and risk, after which we can set

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

In general, such revelations of w and W are laborious. The observed choices depend not only on w and W but also on the utility function U, so that complex multi-parameter estimations must be carried out (Tversky & Kahneman 1992, p. 311).

A second way to elicit B(E) is by measuring the *canonical probability* p of event E, defined through the equivalence

591 $(p{:}x, y) \sim (E{:}x, y)$ (5.2)

592 for some preset $x > y$, say $x = 100$ and $y = 0$. Then $w(B(E))(U(x)-U(y)) = w(p)(U(x)-U(y))$,
593 and $B(E) = p$ follows. Wakker (2004) discussed the interpretation of Eqs. 5.1 and 5.2 as
594 belief.
595 Canonical probabilities were commonly used in early decision analysis (Raiffa 1968,
596 Section 5.3; Yates 1990 pp. 25-27) under the assumption of expected utility. A recent
597 experimental measurement is in Holt (2005, Chapter 30), who also assumed expected utility.
598 A practical difficulty is that the measurement of canonical probabilities requires the
599 measurement of indifferences, and these are not easily inferred from choice. For example,
600 Holt (2005) used the Becker-deGroot-Marschak mechanism, discussed above. Huck &
601 Weiszäcker (2002) compared the QSR to the measurement of canonical probabilities and
602 found that the former is more accurate.
603 A third way to correct reported probabilities is to collect, for each r_E, many events to
604 which the agent assigned the same r_E value in the past and of which we now know whether or
605 not they obtained. We then determine the relative frequency of these events, assuming that
606 this can be taken as the true, objective, probability. We, thus, turn these events into
607 unambiguous events. The distance between r_E and this relative frequency is an index of the
608 (mis)calibration of the agent. This calibration technique has been studied in theoretical game
609 theory (Sandroni, Smorodinsky, & Vohra 2003), and has been applied to weather forecasters
610 (Murphy & Winkler 1974). It needs extensive data, which is especially difficult to obtain for
611 rare events such as earthquakes. It needs further assumptions about the stability of
612 distortions over time, and is hard to apply in experiments of limited durations. These
613 drawbacks were pointed out by Clemen & Lichtendahl (2002), who proposed correction
614 techniques for probability estimates in the spirit of our paper, but still based these on
615 traditional calibration techniques. Our correction ("calibration") technique is considerably
616 more efficient than traditional ones.
617 We now introduce risk corrections that combine the advantages of measuring $B(E) =$
618 $w^{-1}(W(E))$, of measuring canonical probabilities, and of calibrating reported probabilities
619 relative to objective probabilities, while avoiding the problems described above, by
620 benefiting from the efficiency of proper scoring rules. The QSR does entail a restriction of
621 the observations regarding B(E) to cases of E being more likely than E^c (Observation 4.2.1).
622 Note that the right-hand sides of Eqs. 4.2.5 and 4.3.6 are identical. Hence, if we find a p
623 with the same r value as E, then we can, because of Eq. 4.2.5, immediately substitute p for

the right-hand side of Eq. 4.3.6, getting B(E) = p without need to know the ingredients w and U of Eq. 4.3.6. This observation (to be combined with Eq. 2.2 for $r < 0.5$) implies the following corollary, which we display for its empirical importance.

COROLLARY 5.1. Under Eq. 4.3.1, for the optimal choice $r = r_E$, assume that $r > 0.5$. Then

$$B(E) = R^{-1}(r). \tag{5.3}$$

□

This corollary is useful for empirical purposes. It is the only implication of our theoretical analysis that is needed for applications. We first infer the (for the participant) optimal R(p) for a set of exogenously given probabilities p that is so dense (all values $p = j/20$ for $j \geq 10$ in our experiment) that we obtain a sufficiently accurate estimation of R and R^{-1}. Then, for all uncertain events E more likely than their complement, we immediately derive B(E) from the observed r_E through Eq. 5.3. Summarizing:

If for event E the participant reports probability $r_E = r$

and for objective probability p the participant also reports probability $R(p) = r$

then $B(E) = p$.

We, therefore, directly measure the curve R(p) in Figure 3.1 empirically, and apply its inverse to r_E. 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 3.1.

We call the function R^{-1} the *risk-correction* (for proper scoring rules), and $R^{-1}(r_E)$ the *risk-corrected probability*. This value is the canonical probability, obtained without having measured indifferences such as through the Becker-DeGroot-Marschak mechanism, without having measured U and w as in decision theory, and without having measured relative frequencies in many repeated observations of past events with the same reported probabilities as in calibrations. Obviously, if R(p) does not deviate much from p, then no risk correction is needed. Then reported probabilities r directly reflect beliefs, and we have ensured that traditional analyses of QSRs give proper results.

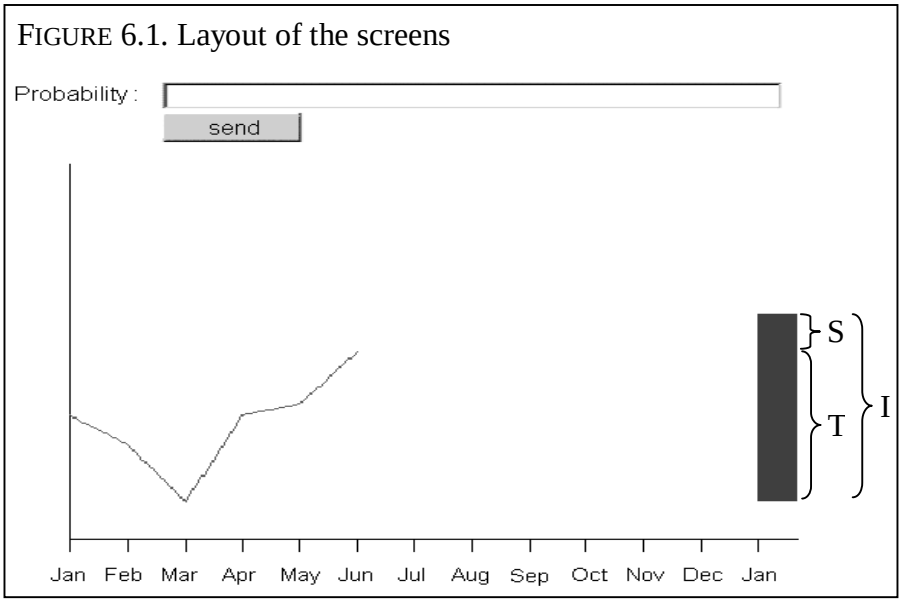
The curves in Figure 3.1 can be reinterpreted as inverses of risk corrections. The examples illustrated there were based on risk averse decision attitudes, leading to conservative estimations moved in the direction of 0.5. Risk seeking will lead to the opposite effect, and will generate overly extreme reported probabilities, suggesting overconfidence. Obviously, if factors in the probability elicitation of the calibration part induce

overconfidence and risk seeking, then our risk correction will detect those biases and correct for them. If, after the risk correction, overconfidence is (still) present, then it cannot be due to risk seeking. It then is convincing that overconfidence is a genuine property of belief, irrespective of risk seeking.

6. An Illustration of Our Measurement of Belief

This section describes risk corrections for a participant in the experiment so as to illustrate how our method can be applied empirically. It will illustrate that the conclusion in Corollary 5.1 is all of the theoretical analysis that is needed for applying 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



The left side of Figure 6.1 displays the performance of stock 20 in our experiment from January 1 until June 1 1991 as given to the participants. It concerned CSM certificates dealing in sugar and bakery-ingredients. Further details (such as the absence of a unit on the y-axis) will be explained in Section 7. The right side of the figure displays two disjoint intervals S and T, and their union $I = S \cup T$. 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 these three questions). For participant 25, the results are as follows.

$r_S = 0.10$; $r_T = 0.55$; $r_I = 0.75$. (6.1)

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

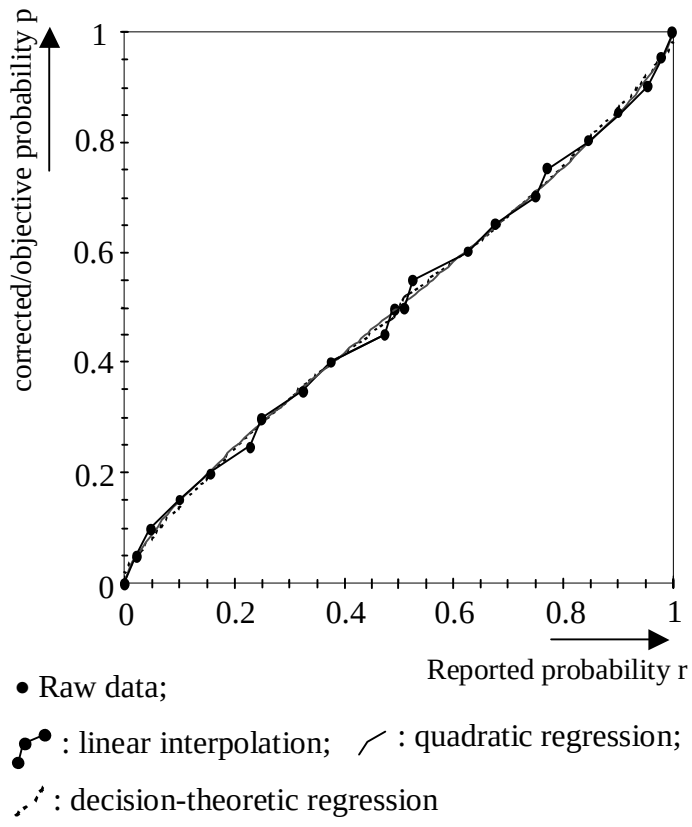
The additivity bias is $0.10 + 0.55 - 0.75 = -0.10$. (6.2)

Table 6.1 and Figure 6.2 (in inverted form) display the reported probabilities R(p) that we measured from this participant, with the points in the figure indicating raw data, and the curves explained later.

TABLE 6.1. Reported probabilities R(p) for given probabilities p of participant 25

p	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
R(p)	.02	.04	.10	.15	.23	.25	.32	.37	.47	.51	.53	.63	.68	.75	.77	.85	.90	.96	.98

FIGURE 6.2. $R^{-1}(r)$ of participant 25



For simplicity of presentation, we analyze the data here using linear interpolation. Then R(0.56) is obtained from

$0.56 = 0.8 \times 0.55 + 0.2 \times 0.60$, R(0.55) = 0.53, and R(0.60) = 0.63,

through

$$R(0.56) = R\big(0.8\times0.55 + 0.2\times0.60\big) = 0.8\times R(0.55) + 0.2\times R(0.60) =$$
$$0.8\times0.53 + 0.2\times0.63 = 0.55.$$

Using this value for R(0.56), further the values R(0.15) = 0.10 and R(0.70) = 0.75, and, finally, Eq. 5.3, we obtain the following risk-corrected beliefs.

$$B(S) = R^{-1}(0.10) = 0.15;\ B(T) = R^{-1}(0.55) = 0.56;\ B(I) = R^{-1}(0.75) = 0.70;$$
the additivity bias is $0.15 + 0.56 - 0.70 = 0.01$. (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 so for repeated payments and decisions (treatment "t=ALL").

It is statistically preferable to fit data with smoother curves than resulting from linear interpolation. We derived "decision-theoretic" parametric curves for R(p) from Corollary 4.2.5, with further assumptions explained at the end of Subsection 8.1.[7] The resulting curve for participant 25 is given in the figure. $B = R^{-1}(r)$ and this curve lead to

$B(S) = R^{-1}(0.10) = 0.15;\ B(T) = R^{-1}(0.55) = 0.54;\ B(I) = R^{-1}(0.75) = 0.71$; the additivity bias is $0.15 + 0.54 - 0.71 = -0.02$, (6.4)

again reducing the uncorrected additivity bias. The quadratic curve gave virtually the same results, and will be discussed in Section 10.1.

---

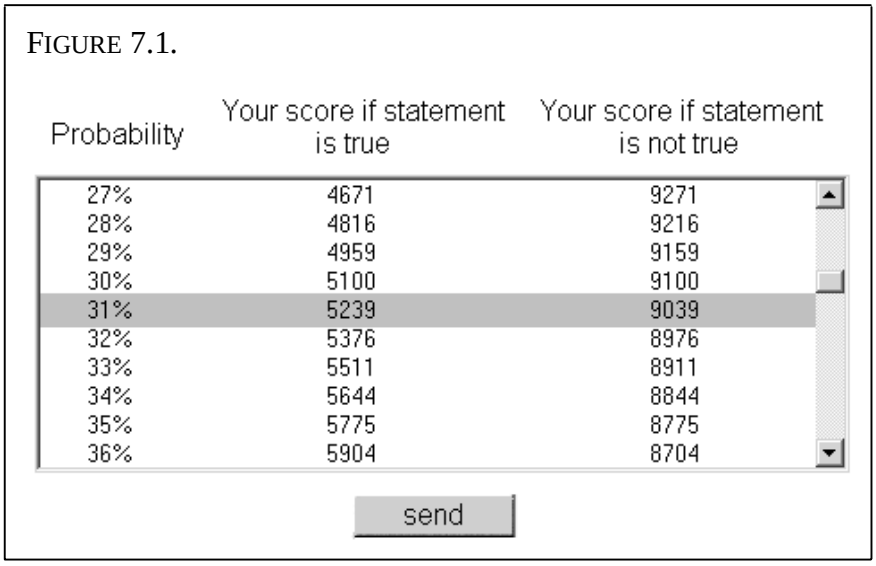
[7] The decision-theoretic curve in the figure is the function $p = B(E) = \dfrac{r}{r + (1-r)\dfrac{1.27\big(1-(1-r)^2\big)^{0.27}}{1.27(1-r^2)^{0.27}}}$, in agreement with Corollaries 5.1 and 4.2.5, where we estimated w(p) = p and found $\rho = 1.27$ as optimal value for U(x) in Eq. 7.1.

7. An Experimental Application of Risk Corrections: Method

Participants. 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 C. 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 (not) true, illustrated in Figure 7.1.

FIGURE 7.1.



All figures (including Figure 6.1) are reproduced here in black and white; in the experiment we used colors 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 an "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. 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 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 two small and one big interval within one stock were not randomized stochastically, but were 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, both moderately small, or one very small and one moderately small.

In the *calibration part* of the experiment, participants made essentially the same decisions as in the stock-price part, but now for 20 events with objective probabilities. Thus, participants simply made choices between risky prospects with objective probabilities. We used two 10-sided dice to determine the outcome of the different prospects and 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 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 decision screen was very similar to Figure 7.1, except for the fact that we wrote "row-percentage" instead of "probability" and "your score if the roll of the die is 01-25" instead of "your score if statement is true;" etc.

Motivating participants. Depending on whether the uncertain event obtained or not and on the reported probability for the uncertain event, a number of points was determined for each

question through the QSR (Eq. 2.1), using 10000 points as 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 through 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 twenty separate compartments containing pairs of 10-sided dice to determine the outcome of each of the twenty prospects at the same time for the treatment t=ALL.

In treatment t=ONE, the random-lottery incentive system was used. That is, at the end of the experiment, one 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 through 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 Observation 4.2.2. Indeed, by Eq. 3.2, 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 the general model of Eq. 4.2.1. We used parametric fittings, where 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 considered for w(p) is Prelec's (1998) popular two-parameter family

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

834 $$w(p) = \left(exp(-\beta(-ln(p))^{\alpha})\right). \qquad (7.2)$$

835 We will mostly use the one-parameter subfamily with $\beta=1$, as in Eq. 4.2.4, for reasons
836 explained later. Substituting the above functions yields

837 $$B(E) = \exp\left(-\left(\frac{-ln\left(\frac{r(2r-r^2)^{1-\rho}}{(1-r)(1-r^2)^{1-\rho} + r(2r-r^2)^{1-\rho}}\right)}{\beta}\right)^{1/\alpha}\right).$$

838 for Eq. 4.3.6.

839 The model we estimate is as follows.

840 $$R_{s,t,k}(j/20) = h(j/20,\alpha_t, \rho_t) + \varepsilon_{s,t,k}(j/20,\sigma_t^2). \qquad (7.3)$$

841 Here $R_{s,t,k}(j/20)$ is the reported probability of participant s for known probability $p=j/20$ ($10 \le$
842 $j \le 19$) in treatment t (t = ALL or t = ONE) for the k^{th} measurement for this probability, with
843 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_t
844 is the remaining probability-weighting parameter (Eq. 7.2), and ρ_t is the power of utility (Eq.
845 7.1). The function h is the inverse of Eq. 4.2.5. Although we have no analytic expression for
846 this inverse, we could calculate it numerically in the analyses. The error terms $\varepsilon_{s,t,k}(j/20)$ are
847 drawn from a truncated normal distribution with mean 0 and treatment dependent variance
848 σ_t^2. The distribution of the error terms is truncated because reported probabilities below 0
849 and above 1 are excluded by design. Error terms are identically and independently
850 distributed across participants and choices. We employ maximum likelihood to estimate the
851 parameters of Eq. 7.3. We also carried out an analysis at the individual level of the
852 calibration part, with $\alpha_{s,t}$ and $\rho_{s,t}$ instead of α_t and ρ_t, i.e. with these parameters depending on
853 the participant.

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

857 $$r_S + r_T = r_I. \qquad (7.4)$$

858 Hence, $r_S + r_T - r_I$ is an index of deviation from additivity, which we call the *additivity bias*
859 of r. For the special case of S the universal event with r a decision-weighting function, Dow

& Werlang (1992) interpreted this quantitative index of nonadditivity as an index of uncertainty aversion.

Under the null hypothesis of additivity for risk-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

$$B(S) + B(T) = B(I). \tag{7.5}$$

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 additivity bias was enlarged or reduced by correction; (c) whether the absolute value of the additivity bias 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 is larger than 0.2 and, hence, 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. 7.3 for β=1 (Eq. 4.2.4 instead of Eq. 7.2) together with the estimates of some models with

additional restrictions. We first give results for group averages, assuming homogeneous participants.

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	NA	11.16* (0.30)	0.91* (0.06)	0.89* (0.14)	10.63* (0.30)	0.85* (0.04)	1.41* (0.07)	6513.84
2	$\alpha_{ONE} = \alpha_{ALL} = \rho_{ONE} = \rho_{ALL} = 1$	12.14* (0.31)	–	–	10.30* (0.26)	–	–	6554.55
3	$\alpha_{ONE} = \rho_{ONE} = 1$	12.14* (0.31)	–	–	10.63* (0.30)	0.85* (0.04)	1.41* (0.07)	6539.04
4	$\alpha_{ALL} = \rho_{ALL} = 1$	11.16* (0.30)	0.91* (0.06)	0.89* (0.14)	10.30* (0.27)	–	–	6529.36
5	$\alpha_{ONE} = \alpha_{ALL}$	11.21* (0.30)	0.87* (0.03)	0.99* (0.08)	10.60* (0.29)	–	1.37* (0.06)	6514.31
6	$\rho_{ONE} = \rho_{ALL}$	11.40* (0.31)	0.79* (0.03)	1.19* (0.07)	10.47* (0.28)	0.96* (0.04)	–	6520.51
7	$\alpha_{ONE} = \alpha_{ALL} = 1$	11.12* (0.29)	–	0.70* (0.04)	10.52* (0.29)	–	1.14* (0.03)	6519.68
8	$\rho_{ONE} = \rho_{ALL} = 1$	11.23* (0.29)	0.87* (0.02)	–	10.43* (0.28)	1.07* (0.02)	–	6522.46
9	$\alpha_{ONE} = \alpha_{ALL} = \rho_{ONE} = 1$	12.14* (0.31)	–	–	10.52* (0.29)	–	1.14* (0.03)	6544.09
10	$\alpha_{ONE} = \alpha_{ALL} = \rho_{ALL} = 1$	11.12* (0.29)	–	0.70* (0.04)	10.30* (0.27)	–	–	6530.14
11	$\alpha_{ONE} = \alpha_{ALL} = 1$, $\rho_{ONE} = \rho_{ALL}$	12.05* (0.34)	–	0.98* (0.03)	10.30* (0.27)	–	–	6554.33

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

Overall need for risk-correction. The 1st row of Table 8.1 shows the results for the most general model. The 2nd row presents the results without any correction. The likelihood reduces significantly (Likelihood Ratio test, p = 0.01) and substantially, so that risk-correction is called for. Risk-correction is also called for in both treatments in isolation, as the 3rd and 4th rows show, which significantly improve the likelihood relative to the 2nd row (Likelihood Ratio test; p = 0.01 for t=ALL, comparing 3rd to 2nd row; p = 0.01 for t=ONE, comparing 4th to 2nd row).

Comparing the two treatments. The likelihood for correcting only t=ALL (3rd row) is worse than for correcting only t=ONE (4th row), suggesting that there is more need for risk-correction for treatment t=ONE than for t=ALL. This difference does not seem to be caused by different probability weighting. The coefficients for probability weighting (α_{ONE}, α_{ALL}) in the 1st row are close to each other and are both smaller than 1. Apparently, probability weighting does not differ between t=ONE and t=ALL. Indeed, adding the restriction $\alpha_{ONE} = \alpha_{ALL}$ (5th row) does not decrease the likelihood of the data significantly (Likelihood ratio test; $p > 0.05$).

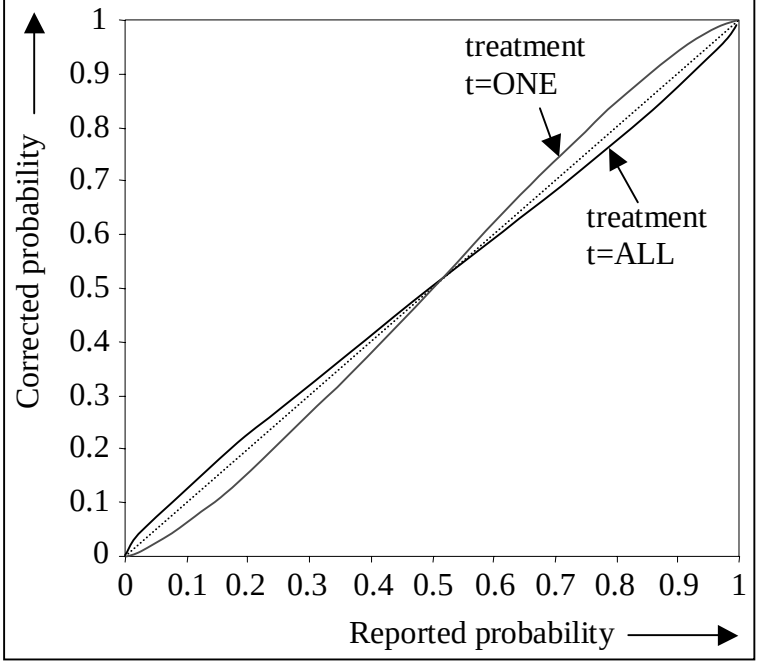
The difference between the two treatments is apparently caused by curvature of utility, captured by ρ_{ONE} and ρ_{ALL}. We obtain $\rho_{ONE} < \rho_{ALL}$: when only one decision is paid out then participants exhibit more concave curvature of utility than when all decisions are paid out. Given same 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 1st row. This restriction significantly reduces the likelihood of observing the data (Likelihood Ratio test, $p = 0.01$).

Comparing utility and probability weighting. Correcting only for utility curvature (7th row, $\alpha_{ONE} = \alpha_{ALL} = 1$) has a somewhat better likelihood than correcting only for probability weighting (8th row, $\rho_{ONE} = \rho_{ALL} = 1$).

Discussion of comparison of utility curvature and probability weighting for group-averages. In deterministic choice, α could be determined through the flat part of R around 0.5, after which ρ could serve to improve the fit elsewhere. Statistically, however, α and ρ have much overlap, with risk aversion enhanced and R(p) moved towards 0.5 by increasing α and decreasing ρ, and one does not add much explanatory power to the other. It is, therefore, better to use only one of these parameters. Another reason to use only one concerns the individual analysis reported in the following subsection. Because we only have 20 choices per participant it is important to economize on the number of free parameters there.

We found above that ρ has a slightly better explanatory power than α. For this reason, and for reasons of convenience (see discussion section), we will only use the parameter ρ, and assume $\alpha = 1$ henceforth. Figure 8.1 displays the resulting average risk-correction for the two treatments separately.

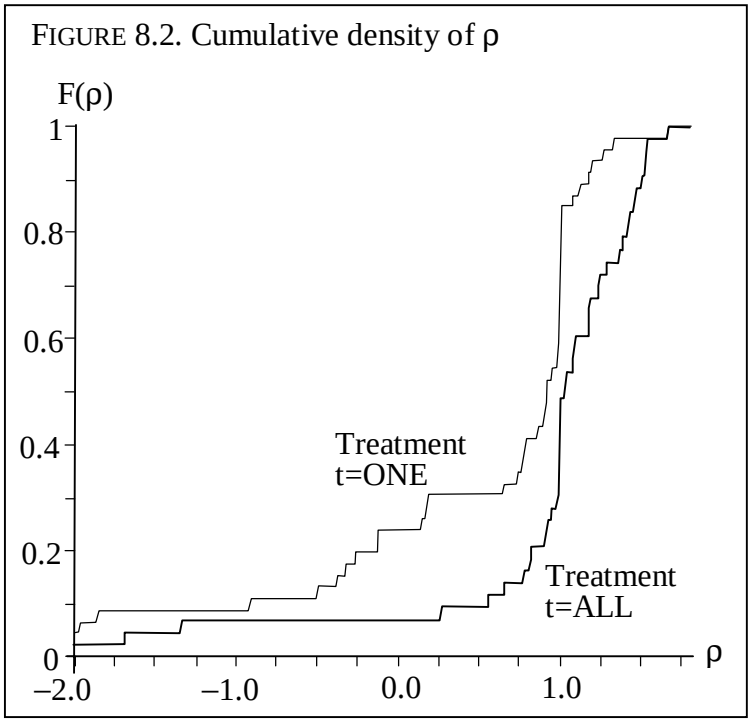
FIGURE 8.1. Corrected versus reported probability



Comparing the two treatments when there is no probability weighting. The average effect of correction for utility curvature is not strong, especially for t=ALL. Yet this correction has a significant effect, as can be seen from comparing the 7th row (general ρ) in Table 8.1 to its 9th row ($\rho_{ALL} = 1$) (Likelihood Ratio test, $p = 0.01$).

8.2. Individual Analyses

Need for risk-correction at the individual level. There is considerable heterogeneity in each treatment. Whereas the corrections required were significant but 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 $\alpha = \beta = 1$. There are wide deviations from the value $\rho=1$ (i.e., no correction) on both sides. As seen from the group-average analysis, there are more deviations at the risk-averse side of $\rho < 1$.

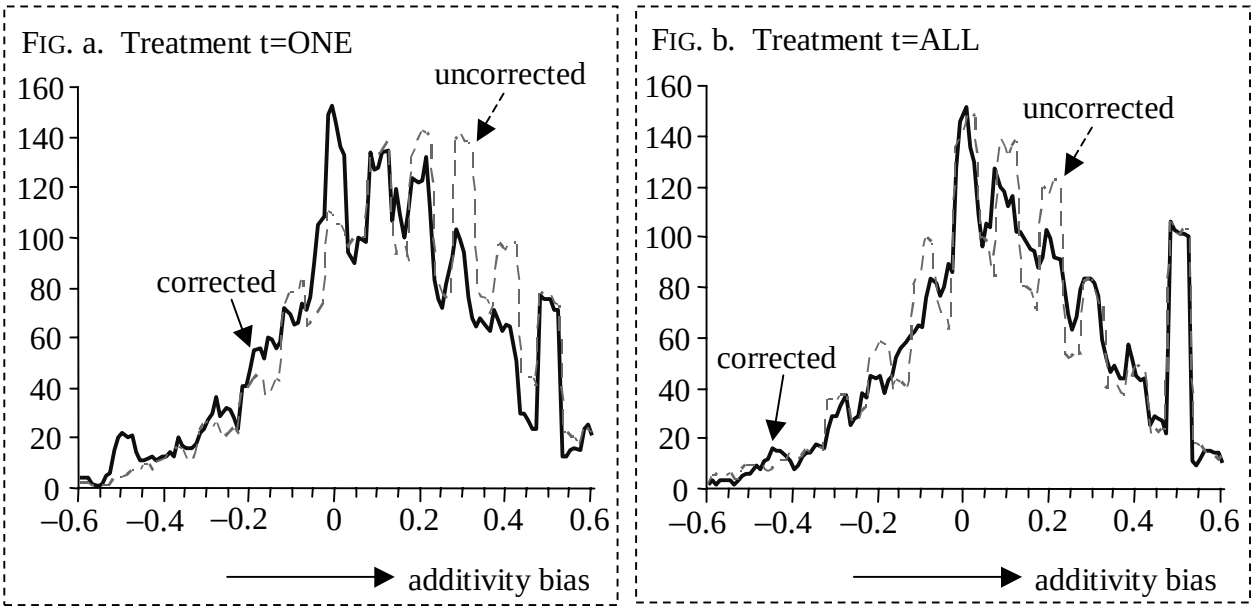


Comparing the two treatments. The ρ-coefficient distribution of treatment t=ONE dominates the ρ-coefficient distribution of treatment t=ALL. A two-sided Mann-Whitney test rejects the null-hypothesis that the ranks of ρ-coefficients are equal across the treatments in favor of the hypothesis that the ρ-coefficients for t=ONE are lower than for t=ALL (p=0.001). It confirms that for group averages there is more risk aversion, 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; etc.

All comparisons hereafter are based on two-sided Wilcoxon signed rank sum tests. Figure 9.1 displays 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. 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.

We first consider 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 per individual, 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 interesting. 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 per individual, 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 precisely, 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 for each individual the average additivity bias over all stocks, 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 stock, where taking for each individual the average additivity bias over all stocks (average 0.237 uncorrected and average 0.239 corrected; $z = -0.36$, $p = 0.7$) also gave no significant difference.

Classifications of individuals according to whether they exhibited more positive or more negative additivity biases, and to whether risk corrections improved or worsened the additivity bias more often, confirmed the patterns obtained above through stockwise analyses, and will not be reported.

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. Discussion

10.1. Discussion of 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 (Section 3). We, therefore, used an alternative explanation that relates properness for unique events to observed frequencies of repeated events (Appendix C).

Besides the family of Prelec (1998), other parametric families for weighting functions have been used in the literature, such as the family of Tversky & Kahneman (1992), and the one of Goldstein & Einhorn (1987, Eqs. 22–24). We used Prelec's family because it performs equally well empirically as the other families but is analytically more tractable, for example because its inverse can be defined easily (cf. Corollary 4.3.4). In addition, it is the only one having an axiomatic foundation.

In pragmatic applications of our method, 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. The curve is virtually indistinguishable from the decision-theoretic curve. This observation, together with Corollary 5.1 demonstrating 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 easy. The theoretical part 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 risk-corrected reported probabilities, then one may choose the pragmatic shortcuts just described.

10.2. Discussion of Main Results

The significantly positive additivity bias that we found in all analyses shows that the separate intervals together receive more weight than their union. This finding agrees with other empirical findings in the literature, and it underlies the subadditivity of support theory (Tversky & Koehler 1994).

After some theoretical debates about the random-lottery incentive system (Holt 1986), as in our treatment t=ONE, the system was tested empirically and found to be incentive-compatible (Starmer & Sugden 1991). It is today the almost exclusively used incentive system for measurements of individual preferences (Holt & Laury 2002). Unlike repeated payments it avoids income effects such as Thaler & Johnson's (1990) house money effect, and the drift towards expected value and linear utility. 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.

For some applications group averages of probability estimates are most relevant, such as when aggregating expert judgments or predicting group behavior. 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.

10.3. Discussion of Further Results

The lack of extra explanatory power of parameter β in Eq. 7.2 should come as no surprise because β and α behave similarly 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 4.2.2).

We found that the risk correction through the utility curvature parameter ρ fitted the data somewhat better than the correction through the probability-weighting parameter α. This finding may be interpreted as some descriptive support for expected utility. Another reason that we used ρ and not α in our main analysis is that ρ, and utility curvature, are more well-known in the economic literature than probability weighting, and are more analytically tractable with R^{-1} defined everywhere. Although ρ indeed reflects the power of utility *if expected utility is assumed*, we caution against unqualified interpretations here, as in any study of risk aversion. The parameter ρ may also capture risk aversion generated by probability weighting, and possibly by other factors.

10.4. General Discussion

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. 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.

A drawback of our risk-correction procedure is that it requires individual measurements of QSRs for given probabilities. If it is not possible to obtain individual measurements, 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 good as is possible without requiring extra measurements. In this respect the average curves in our Figure 8.1 are reassuring for existing studies, because these curves suggest that only small corrections were called for regarding the group averages in our context.

Allen (1988) proposed to avoid biases of the QSR due to nonlinear utility by paying in the probability of winning a prize instead of paying in money, and this procedure was implemented by McKelvey & Page (1990). The procedure, however, only works if expected utility holds, and there is much evidence against this assumption. Indeed, Selten, Sadrieh, &

Abbink (1999) showed empirically that payment in probability does not generate the desired risk neutral behavior.

11. Conclusion

Applications of proper scoring rules to measure subjective beliefs have so far been based on the assumption of expected value maximization. However, many empirical deviations of this assumption have been found. We have provided a method to correct for such deviations, and have proved theoretically that our method provides such corrections under the modern theories of nonexpected utility. These theories are empirically more realistic than expected value maximization.

We have demonstrated the feasibility and empirical tractability of our method in an experiment, where we used it to investigate some properties of quadratic proper scoring rules and beliefs. In a treatment with one big incentive for one randomly selected decision, we found systematic distortions of (uncorrected) reported probabilities. No systematic distortions were found in a treatment with many repeated decisions and repeated small payments. When applying our correction procedure, both treatments give similar deviations from additivity for the (corrected) beliefs. This finding suggests that subjective beliefs are genuinely nonadditive. It means that expected value and expected utility are violated at the level of beliefs, and that beliefs and ambiguity attitudes cannot be expressed in terms of traditional probabilities, in agreement with Ellsberg's demonstrations. More general nonadditive measures, such as used in nonexpected utility theories, are called for. For belief elicitations where no risk correction can be implemented, repeated decisions with repeated small payments are preferable to single large payments.

Appendix A. Proofs and Technical Remarks

In Eqs. 4.2.1 and 4.2.2, probability p has a different decision weight when it yields the best outcome of the prospect ($r > 0.5$) than when it yields the worst ($r < 0.5$). Similarly, in Eqs. 4.3.1 and 4.3.2, E has a different decision weight when it yields the highest outcome ($r > 0.5$) than when it yields the lowest outcome ($r < 0.5$). Such a dependency of decision weights on the ranking position of the outcome is called *rank-dependence* in the literature.

Under rank-dependence, the sum of the decision weights in the evaluation of a prospect are 1 even though w(B(E)) is not additive in E. This property is necessary for the functional that evaluates prospects to satisfy natural conditions such as stochastic dominance, which explains why theoretically sound nonexpected utility models could only be developed after the discovery of rank dependence, a discovery that was made independently by Quiggin (1982) for the special case of risk and by Schmeidler (1989, first version 1982) for the general context of uncertainty.

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 $r = 1$ hereafter can be treated as interior solutions, and they satisfy the first-order optimality conditions.

PROOF OF OBSERVATION 4.1.3. If $r = 0.5$ then the marginal utility ratio in Eq. 4.1.2 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. 4.1.2 would be at least 1 so that the right-hand side of Eq. 4.1.2 would at most be 0.5, contradiction $r > 0.5$. 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. 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 THEOREM 4.3.2. We write π for the decision weight W(E). For optimality of interior solutions r, the first-order optimality condition for Eq. 4.3.1 is that

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

implying

$$\pi(1-r)U'(a-b(1-r)^2) = (1-\pi)rU'(a-br^2) \qquad (A.1)$$

or $\pi U'(a-b(1-r)^2) = r \times (\pi U'(a-b(1-r)^2) + (1-\pi)U'(a-br^2))$, and Eq. 4.3.5 follows.

□

PROOF OF COROLLARY 4.3.4. Let $r > 0.5$ be optimal, and write $\pi = W(E)$. Then Eq. A.1 implies

$\pi \times \left((1-r)U'(a-b(1-r)^2) + rU'(a-br^2)\right) = rU'(a-br^2)$, implying

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

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 4.3.4 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 5.1. Theorem 4.3.2 implies that the right-hand side of Eq. 4.3.5 is r both as is, and with p substituted for B(E). Because Eq. 4.3.5 is strictly increasing in w(B(E)), and w is strictly increasing too, p = B(E) follows. □

Appendix B. Models for Decision under Risk and Uncertainty

For binary (two-outcome) prospects with both outcomes nonnegative, as considered in QSRs, Eqs. 4.3.1 and 4.3.2 have appeared many times in the literature. Early references include Allais (1953, Eq. 19.1), Edwards (1954 Figure 3), and Mosteller & Nogee (1951, p. 398). The convenient feature that binary prospects suffice to identify utility U and the nonadditive $w \circ B = W$ was pointed out by Ghirardato & Marinacci (2001), Gonzalez & Wu (2003), Luce (1991, 2000), Miyamoto (1988), Pfanzagl (1959, p. 287 top and middle), and Wakker & Deneffe (1996, p. 1143 and pp.1144-1145). Luce & Narens (1985, Theorems 7.1 & 7.2.2) and Ghirardato, Maccheroni, & Marinacci (2005) showed that the form in Eqs. 4.3.1 and 4.3.2 is essentially the only one for binary prospects that allows for interval-scale utility of outcomes.

The convenient feature that virtually all existing decision theories agree on the evaluation of binary prospects was pointed out by Miyamoto (1988), calling Eqs. 4.3.1 and 4.3.2 generic utility, and Luce (1991), calling these equations binary rank-dependent utility. It was most clearly analyzed by Ghirardato & 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, Chapter 3) and in Gonzalez & Wu (2003). Only for more than two outcomes, the theories diverge (Mosteller & Nogee 1951 p. 398; Luce 2000, introductions to Chapters 3 and 5). We next describe some of the mentioned decision theories. Because we consider only nonnegative outcomes, losses play no role, and we describe prospect theory only for gains hereafter.

We begin with decision under risk, with known objective probabilities P(E). Expected utility (von Neumann & Morgenstern, 1944) is the special case where w is the identity and B(E) = P(E). Kahneman & Tversky's (1979) original prospect theory, Quiggin's (1982) rank-dependent utility, and Tversky & 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 (Savage 1954) 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 & 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 & Schmeidler 1989, Wald 1950) results with W(E) the infimum value P(E) over all priors P. Under Machina & Schmeidler's (1992) probabilistic sophistication, B is an additive probability measure.

Appendix C. Experimental Instructions

[Instructions are translated from Dutch and concern the instructions of Treatment t=ONE only]

This experiment is about statements of which you do not know whether they are true or not. An example is the statement that snow did fall in Amsterdam in March 1861. You do not know for sure whether this statement is true or not. We will ask you to indicate how likely it is for you that such a statement is true, using probability judgments expressed in percentages. Perhaps you will, for example, attach a probability of 30% to the statement that it snowed in March 1861 in Amsterdam. We will then determine a score for you with the help of the added table *on paper*.

According to the table, for a probability judgment of 30% you get score 5100 if the statement is true (snow did fall in Amsterdam in March 1861). You get score 9100 if the statement is not true (snow did not fall in Amsterdam in March 1861). If you give a different probability judgment, you get different scores, as shown in the table. For example, if you give a probability judgment of 100%, your score is 10000 if the statement is true (snow did fall), and 0 if the statement is not true (snow did not fall). We now like to check whether the table with the scores is clear.

[Practice questions using the table]

Your answers were right. We will now explain some further features of the table. If you are certain that the statement is true, then it is best for you to give the maximum probability judgment of 100% because that gives the maximum score 10000 for a true statement. Every other answer than surely yields a lower score. If you are certain that the statement is not true, then it is similarly best to give the minimum probability judgment of 0%, because that gives the maximum score 10000 for a false statement. In many cases you do not know for certain whether a statement is true or not. We will now explain an important feature of the table on the basis of a thought experiment.

Thought experiment about repeated statements

The properties of the table can be well illustrated with the help of repeated statements. Imagine, as a thought experiment, that you first have to give your probability judgment about a particular statement (for example, snow in Amsterdam in a particular year, say 1861). Imagine that you give judgment 30%, which means that you earn 5100 points in case of snow and 9100 points in case of no snow. Next however, various repetitions of that statement are being considered (snow in Amsterdam in March 1862, snow in Amsterdam in March 1863, …., snow in Amsterdam in March 1960), leading to a total of 100 of such statements. For all 100 statements (thus every year between 1861 and 1960) your score will be determined according to the table and your probability judgment (that is the same for every 100 statements). Your total score is then equal the sum of those 100 scores. For example, if it did snow in Amsterdam in March 35 times in those 100 years, and it did not snow 65 times, a probability judgment of 30% yields the following total score:

35 x 5100 + 65 x 9100 = 770000

We can also calculate this for other probability judgments, suppose that your probability judgment was 35%, then your total score was:

35 x 5775 + 65 x 8775 = 772500

On the next page we show that your total-score is optimal if your probability judgment is exactly equal to that percentage. Put differently, if for example 35 of the 100 (35%) statements are true, then it is best for you to choose probability judgment 35% because it will give you the highest total-score.

Now suppose that 35 of the 100 statements are true

We will determine what your total-score would have been at different judgments.

[Table showing the total score for all possible probability judgments]

It looks like judgment 35 is best. We conclude that if 35% of the statements are true, probability judgment 35 is optimal. Something similar holds for every percentage.

CONCLUSION. For every percentage of true statements your total-score is optimal if you choose your probability judgment to be equal to that percentage. Check this for another number by clicking on continue.

Now suppose that [entered number] of the 100 statements are true

We will determine what your total-score would have been at different judgments.

[Table showing the total score for all possible probability judgments]

It looks like judgment [entered number] is best. Thus we conclude that for [entered number] % true statements, probability judgment [entered number] is optimal. Something similar holds for every percentage.

CONFIRMATION OF THE CONCLUSION. For every percentage of true statements your total-score is optimal if you choose your probability judgment to be equal to that percentage. If you want, you can check the conclusion again for another number than [entered number] by clicking on the link below.

The experiment for non-repeated statements

The experiment we will perform concerns unique, and not repeated, statements. The various unique statements we consider are all different. For every single one of them you can give a different probability judgment.

There is a big difference between the real experiment and the thought-experiment with repetition. In the thought experiment there was an objective-optimal probability judgment, based on the percentage of true statements. In the real experiment, there are no repetitions and for every probability judgment you get only one score.

The thought experiment does give a guide for your probability judgment in the real experiment, with the percentage true statements as reference point. It is now based on your own subjective judgment however, and not on objective calculations. In the real experiment, there is no right or wrong answer. You purely choose what you like best.

In the experiment, you will encounter all different sorts of statements, more or less probable ones, and you can choose all probability judgments ranging from 0% till 100%. You can only choose whole percentages.

Payoff

This experiment consists of two parts. In both parts you will be asked to give probability judgments, 100 in part 1 and 20 in part 2. At the end of the experiment, one out of 120 statements considered during the experiment will be randomly (with equal probability) selected and on the basis of your score at this statement you will be paid out in euros, where 500 points is equal to 1 euro. Click on continue to read the instructions of the first part of the experiment.

Instructions part 1

In the graph below you see the price of a stock from January till June in a year in the past. We used real stock prices of the Amsterdam Exchange when we made the graphs. The graph is scaled in such a way that the price of the stock always stays between the upper and lower axis. The same holds for the other graphs you will see later in this experiment. We consider the following statement: on the 31st of December in that particular year, the price of the stock in the graph was in the purple area. We ask you to give a probability judgment about the truth of this statement without any further information about the stock or the year. You can only base this on the course of the graph in the first half of the year.

[Figure showing an example of graph of stock price]

Your score at this question depends on your probability judgment and whether the statement is true or not, according to the table.

[Figure showing the same graph but with three different end prices at 31st of December]

The input of your probability judgment takes place in two phases: first you type in an integer number between 0 and 100, next you will be shown a menu in which your choice is reproduced with the corresponding scores from the table. At that moment you can still alter your choice and choose any other integer between 0 and 100. You can do this by selecting the up or down arrow, or by clicking the mouse in the menu and scroll to another probability

judgment. Next, when you click on OK your choice is final and you continue with the next statement. If you have any questions at this moment, raise your hand. The experimenter will come to you.

Instructions part 2

Part 1 of the experiment is now over. The second part of the experiment consists of 20 statements. Also in this part of the experiment you will be asked to give probability judgments. The difference is that it does not concern the prediction of stock prices now, but rolls with two 10-sided dice. On one of the dice are the values 00, 10, 20, 30, 40, 50, 60, 70, 80, 90 and on the other die are the values 1, 2, 3, 4, 5, 6, 7, 8, 9. Both dice will be rolled. The sum of the outcomes has the values 1-100 (we consider the roll 00-0 as if it is 100), where all values have the same probability.

[Picture showing the two ten sided dice]

An example of a statement is "the outcome is in the range 01-25." This statement is true when the outcome of the dice is indeed between 1 and 25 (including 25), and not true when the outcome is higher than 25. The input of your probability judgment again takes place in two phases: first you type in an integer number between 0 and 100, next you will be shown a menu in which your choice is replicated with the corresponding scores from the table. At that moment you can still alter your choice and choose any other integer number between 0 and 100. You can do this by selecting the up or down arrow, or by clicking the mouse in the menu and scroll to another probability judgment. Next, when you click on OK your choice is final and you continue with the next statement. Also in this part there is no right or wrong answer; you again choose what you want best. At the end of the experiment one statement will be selected and paid out. In case that this is a statement from part 2 of the experiment, you will be asked to roll the two ten sided dice once.

This is the end of part 2. Please raise your hand. The experimenter will come by so that it can be determined which round will be paid out.

References

Abdellaoui, Mohammed (2000), "Parameter-Free Elicitation of Utilities and Probability Weighting Functions," *Management Science* 46, 1497–1512.

Albers, Wulf, Robin Pope, Reinhard Selten, & Bodo Vogt (2000), "Experimental Evidence for Attractions to Chance," *German Economic Review* 1, 113–130.

Allais, Maurice (1953), "Fondements d'une Théorie Positive des Choix Comportant un Risque et Critique des Postulats et Axiomes de l'Ecole Américaine," *Colloques Internationaux du Centre National de la Recherche Scientifique (Econométrie)* 40, 257–332. Paris: Centre National de la Recherche Scientifique. Translated into English, with additions, as "The Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulates and Axioms of the American School," *in* Maurice Allais & Ole Hagen (1979, Eds), *Expected Utility Hypotheses and the Allais Paradox*, 27–145, Reidel, Dordrecht, The Netherlands.

Allen, Franklin (1987), "Discovering Personal Probabilities when Utility Functions are Unknown," *Management Science* 33, 542–544.

Becker, Gordon M., Morris H. de Groot, & Jacob Marschak (1964), "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science* 9, 226–232.

Becker, Selwyn W. & Fred O. Brownson (1964), "What Price Ambiguity? Or the Role of Ambiguity in Decision Making," *Journal of Political Economy* 72, 62–73.

Bernoulli, Daniel (1738), "Specimen Theoriae Novae de Mensura Sortis," *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 5, 175–192. Translated into English by Louise Sommer (1954), "Exposition of a New Theory on the Measurement of Risk," *Econometrica* 22, 23–36.

Reprinted in Alfred N. Page (1968, Ed.), *Utility Theory: A Book of Readings*, Chapter 11, Wiley, New York. Revised translation in William J. Baumol & Stephen M. Goldfeld (Eds, 1968), *Precursors in Mathematical Economics: An Anthology*. Clowes and Sons, London, Selection 2, 15–26.

Bleichrodt, Han & José Luis Pinto (2000), "A Parameter-Free Elicitation of the Probability Weighting Function in Medical Decision Analysis," *Management Science* 46, 1485–1496.

Bliss, Robert R. & Nikolaos Panigirtzoglou (2004), "Option-Implied Risk Aversion Estimates," *Journal of Finance* 59, 407–446.

Braga, Jacinto & Chris Starmer (2005), "Preference Anomalies, Preference Elicitation, and the Discovered Preference Hypothesis," *Environmental and Resource Economics* 32, 55–89.

Brier, Glenn W. (1950), "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review* 78, 1–3.

Camerer, Colin F. & Martin Weber (1992), "Recent Developments in Modelling Preferences: Uncertainty and Ambiguity," *Journal of Risk and Uncertainty* 5, 325–370.

Clemen, Robert T. & Kenneth C. Lichtendahl (2002), "Debiasing Expert Overconfidence: A Bayesian Calibration Model, Fuqua School of Business, Duke University, Durham NC."

Clemen, Robert T. & Fred Rolle (2001), "In Theory … In Practice," *Decision Analysis Newsletter* 20, No 1.

de Finetti, Bruno (1931), "Sul Significato Soggettivo della Probabilità," *Fundamenta Mathematicae* 17, 298–329. Translated into English as "On the Subjective Meaning of Probability," *in* Paola Monari & Daniela Cocchi (Eds, 1993) "P*robabilità e Induzione,"* Clueb, Bologna, 291–321.

de Finetti, Bruno (1937), "La Prévision: Ses Lois Logiques, ses Sources Subjectives," *Annales de l'Institut Henri Poincaré* 7, 1–68. Translated into English by Henry E. Kyburg Jr., "Foresight: Its Logical Laws, its Subjective Sources," *in* Henry E. Kyburg Jr. & Howard E. Smokler (1964, Eds), *Studies in Subjective Probability*, 93–158, Wiley, New York; 2nd edition 1980, 53–118, Krieger, New York.

de Finetti, Bruno (1962), "Does It Make Sense to Speak of "Good Probability Appraisers"?". *In* Isidore J.Good (Ed.), *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, William Heinemann Ltd., London.
Reprinted as Chapter 3 in de Finetti, Bruno (1972), "*Probability, Induction and Statistics."* Wiley, New York.

Dempster, Arthur P. (1967), "Upper and Lower Probabilities Induced by a Multivalued Mapping," *Annals of Mathematical Statistics* 38, 325–339.

Dow, James & Sérgio R.C. Werlang (1992), "Uncertainty Aversion, Risk Aversion and the Optimal Choice of Portfolio," *Econometrica* 60, 197–204.

Echternacht, Gary J. (1972), "The Use of Confidence Testing in Objective Tests," *Review of Educational Research* 42, 217–236.

Edwards, Ward (1954), "The Theory of Decision Making," *Psychological Bulletin* 51, 380–417.

Ellsberg, Daniel (1961), "Risk, Ambiguity and the Savage Axioms," *Quarterly Journal of Economics* 75, 643–669.

Fischer, Gregory W. (1982), "Scoring Rule Feedback and the Overconfidence Syndrome in Subjective Probability Forcasting," *Organizational Behavior and Human Performance* 29, 357–369.

Fischhoff, Baruch, Paul Slovic & Sarah Lichtenstein (1977), "Knowing with Certainty: The Appropriateness of Extreme Confidence," *Journal of Experimental Psychology: Human Perception and Performance* 3, 552–564.

Ghirardato, Paolo, Fabio Maccheroni, & Massimo Marinacci (2005), "Certainty Independence and the Separation of Utility and Beliefs," *Journal of Economic Theory* 120, 129–136.

Ghirardato, Paolo & Massimo Marinacci (2001), "Risk, Ambiguity, and the Separation of Utility and Beliefs," *Mathematics of Operations Research* 26, 864–890.

Gilboa, Itzhak (1987), "Expected Utility with Purely Subjective Non-Additive Probabilities," *Journal of Mathematical Economics* 16, 65–88.

Gilboa, Itzhak & David Schmeidler (1989), "Maxmin Expected Utility with a Non-Unique Prior," *Journal of Mathematical Economics* 18, 141–153.

Gilboa, Itzhak & David Schmeidler (1999), "*A Theory of Case-Based Decisions."* Cambridge University Press, Cambridge, UK.

Goldstein, William M. & Hillel J. Einhorn (1987), "Expression Theory and the Preference Reversal Phenomena," *Psychological Review* 94, 236–254.

Gonzalez, Richard & George Wu (1999), "On the Shape of the Probability Weighting Function," *Cognitive Psychology* 38, 129–166.

Gonzalez, Richard & George Wu (2003), "Composition Rules in Original and Cumulative Prospect Theory,." Graduate School of Business, University of Chicago, Chicago, IL.

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

Guala, Francesco (2000), "Artefacts in Experimental Economics: Preference Reversals and the Becker-DeGroot-Marschak Mechanism," *Economics and Philosophy* 16, 47–75.

Gul, Faruk (1991), "A Theory of Disappointment Aversion," *Econometrica* 59, 667–686.

Hanson, Robin (2002), Piece entitled "Wanna Bet?" in *Nature* 420, November 2002, pp. 354–355.

Holt, Charles A. (1986), "Preference Reversals and the Independence Axiom," *American Economic Review* 76, 508–513.

Holt, Charles A. (2005), "*Webgames and Strategy: Recipes for Interactive Learning*." in press.

Holt, Charles A. & Susan K. Laury (2002), "Risk Aversion and Incentive Effects," *American Economic Review* 92, 1644–1655.

1508 Huck, Steffen & Georg Weiszäcker (2002), "Do Players Correctly Estimate What Others Do?
1509 Evidence of Conservatism in Beliefs," *Journal of Economic Behavior and Organization*
1510 47, 71–85.
1511 Kadane, Joseph B. & Robert L. Winkler (1988), "Separating Probability Elicitation from
1512 Utilities," *Journal of the American Statistical Association* 83, 357–363.
1513 Kahneman, Daniel & Amos Tversky (1979), "Prospect Theory: An Analysis of Decision
1514 under Risk," *Econometrica* 47, 263–291.
1515 Karni, Edi & Zvi Safra (1987), "Preference Reversal and the Observability of Preferences by
1516 Experimental Methods," *Econometrica* 55, 675–685.
1517 Keren, Gideon B. (1991), "Calibration and Probability Judgments: Conceptual and
1518 Methodological Issues," *Acta Psychologica* 77, 217–273.
1519 Keynes, John Maynard (1921), "*A Treatise on Probability.*" McMillan, London. Second
1520 edition 1948.
1521 Knight, Frank H. (1921), "*Risk, Uncertainty, and Profit.*" Houghton Mifflin, New York.
1522 Lehrer, Ehud (2001), "Any Inspection is Manipulable," *Econometrica* 69, 1333–1347.
1523 Liberman, Varda & Amos Tversky (1993), "On the Evaluation of Probability Judgments:
1524 Calibration, Resolution, and Monotonicity," *Psychological Bulletin* 114, 162–173.
1525 Luce, R. Duncan (1991), "Rank- and-Sign Dependent Linear Utility Models for Binary
1526 Gambles," *Journal of Economic Theory* 53, 75–100.
1527 Luce, R. Duncan (2000), "*Utility of Gains and Losses: Measurement-Theoretical and*
1528 *Experimental Approaches.*" Lawrence Erlbaum Publishers, London.
1529 Luce, R. Duncan & Louis Narens (1985), "Classification of Concatenation Measurement
1530 Structures According to Scale Type," *Journal of Mathematical Psychology* 29, 1–72.
1531 Machina, Mark J. (1982), " 'Expected Utility' Analysis without the Independence Axiom,"
1532 *Econometrica* 50, 277–323.
1533 Machina, Mark J. & David Schmeidler (1992), "A More Robust Definition of Subjective
1534 Probability," *Econometrica* 60, 745–780.
1535 Manski, Charles F. (2004), "Measuring Expectations," *Econometrica* 72, 1329–1376.
1536 Marinacci, Massimo (2002), "Probabilistic Sophistication and Multiple Priors,"
1537 *Econometrica* 70, 755–764.
1538 McClelland, Alastair & Fergus Bolger (1994), "The Calibration of Subjective Probabilities:
1539 Theories and Models 1980–1994." *In* George Wright & Peter Ayton (Eds), *Subjective*
1540 *Probability*, 453–481, Wiley, New York.

McKelvey, Richard & Talbot Page (1990), “Public and Private Information: An Experimental Study of Information Pooling,” *Econometrica* 58, 1321–1339.

Miyamoto, John M. (1988), “Generic Utility Theory: Measurement Foundations and Applications in Multiattribute Utility Theory,” *Journal of Mathematical Psychology* 32, 357–404.

Mosteller, Frederick & Philip Nogee (1951), “An Experimental Measurement of Utility,” *Journal of Political Economy* 59, 371–404.

Murphy, Allan H. & Robert L. Winkler (1974), “Subjective Probability Forecasting Experiments in Meteorology: Some Preliminary Results,” *Bulletin of the American Meteorological Society* 55, 1206–1216.

Nyarko, Yaw & Andrew Schotter (2002), “An Experimental Study of Belief Learning Using Elicited Beliefs,” *Econometrica* 70, 971–1005.

Palmer, Tim N. & Renate Hagedorn (2006, Eds), "*Predictability of Weather and Climate.*" Cambridge University Press, Cambridge.

Pfanzagl, Johann (1959), “A General Theory of Measurement —Applications to Utility,” *Naval Research Logistics Quarterly* 6, 283–294.

Plott, Charles R. & Kathryn Zeiler (2005), “The Willingness to Pay-Willingness to Accept Gap, the “Endowment Effect,” Subject Misconceptions, and Experimental Procedures for Eliciting Valuations,” *American Economic Review* 95, 530–545.

Prelec, Drazen (1998), “The Probability Weighting Function,” *Econometrica* 66, 497–527. (First version: Prelec, Drazen (1989), “On the Shape of the Decision Weight Function,” Harvard Business School, Harvard University, Cambridge, MA, USA.)

Prelec, Drazen (2004), “A Bayesian Truth Serum for Subjective Data,” *Science* 306, October 2004, 462–466.

Quiggin, John (1982), “A Theory of Anticipated Utility,” *Journal of Economic Behaviour and Organization* 3, 323–343.

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

Sandroni, Alvaro, Rann Smorodinsky, & Rakesh V. Vohra (2003), “Calibration with Many Checking Rules,” *Mathematics of Operations Research* 28, 141–153.

Savage, Leonard J. (1954), “*The Foundations of Statistics.”* Wiley, New York. (Second edition 1972, Dover Publications, New York.)

Savage, Leonard J. (1971), “Elicitation of Personal Probabilities and Expectations,” *Journal of the American Statistical Association* 66, 783–801.

Schmeidler, David (1989), "Subjective Probability and Expected Utility without Additivity," *Econometrica* 57, 571–587.

Selten, Reinhard, Abdolkarim Sadrieh, & Klaus Abbink (1999), "Money Does not Induce Risk Neutral Behavior, but Binary Lotteries Do even Worse," *Theory and Decision* 46, 211–249.

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

Spiegelhalter, David J. (1986), "Probabilistic Prediction in Patient Management and Clinical Trials," *Statistics in Medicine* 5, 421–433.

Staël von Holstein, Carl-Axel S. (1972), "Probabilistic Forecasting: An Experiment Related to the Stock Market," *Organizational Behaviour and Human Performance* 8, 139–158.

Starmer, Chris & Robert Sugden (1991), "Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation," *American Economic Review* 81, 971–978.

Thaler, Richard H. & Eric 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, Amos & Daniel Kahneman (1992), "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty* 5, 297–323.

Tversky, Amos & Derek J. Koehler (1994), "Support Theory: A Nonextensional Representation of Subjective Probability," *Psychological Review* 101, 547–567.

von Neumann, John & Oskar Morgenstern (1944, 1947, 1953), "*Theory of Games and Economic Behavior."* Princeton University Press, Princeton NJ.

Wakker, Peter P. (2004), "On the Composition of Risk Preference and Belief," *Psychogical Review* 111, 236–241.

Wakker, Peter P. & Daniel Deneffe (1996), "Eliciting von Neumann-Morgenstern Utilities when Probabilities Are Distorted or Unknown," *Management Science* 42, 1131–1150.

Wald, Abraham (1950), "*Statistical Decision Functions."* Wiley, New York.

Winkler, Robert L. (1967), "The Assessment of Prior Distributions in Bayesian Analysis," *Journal of the American Statistical Association* 62, 776–800.

Winkler, Robert L. & Allan H. Murphy (1970), "Nonlinear Utility and the Probability Score," *Journal of Applied Meteorology* 9, 143–148.

Wright, William F. (1988), “Empirical Comparison of Subjective Probability Elicitation Methods,” *Contemporary Accounting* 5, 47–57.

Yates, J. Frank (1990), “*Judgment and Decision Making.”* Prentice Hall, London.