A Truth-Serum for Non-Bayesians: 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

October, 2007

ABSTRACT. Proper scoring rules provide convenient and highly efficient tools for eliciting subjective beliefs. As traditionally used, however, they are valid only under expected value maximization. This paper shows how proper scoring rules can be generalized to modern ("nonexpected utility") theories of risk and ambiguity, yielding mutual benefits: the empirical realism of nonexpected utility is introduced in proper scoring rules, and the beauty and efficiency of proper scoring rules is introduced in nonexpected utility. An experiment demonstrates the feasibility of our generalized proper scoring rule, yielding plausible empirical results.

KEY WORDS: belief measurement, proper scoring rules, ambiguity, Knightian uncertainty, subjective probability, nonexpected utility
JEL-CLASSIFICATION: D81, C60, C91

[1] A preliminary version of this paper circulated with the title "Is the Quadratic Scoring Rule Really Incentive Compatible?" This paper received helpful and detailed comments from Glenn Harrison.

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 and incentive-compatible 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 or indifferences, commonly used in decision theory, because binary choices only give inequalities and approximations, and indifferences are hard to elicit.

The measurement of subjective beliefs is important in many domains (Gilboa & Schmeidler 1999; Machina & Schmeidler 1992; 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), finance (Shiller, Kon-Ya, & Tsutsui 1996), medicine (Spiegelhalter 1986), politics (Tetlock 2005), psychology (McClelland & Bolger 1994), and other fields (Hanson 2002; Johnstone 2006; Prelec 2004). Proper scoring rules are especially useful for giving experts incentives to exactly reveal their degrees of belief. They are commonly used, for instance, to measure the degree of belief of weather forecasters and to improve their calibration (Palmer & Hagedorn 2006; Yates 1990). They have recently become popular in experimental economics and game theory. The quadratic scoring rule is the most popular proper scoring rule today (McKelvey & Page 1990; Nyarko & Schotter 2002; Palfrey & Wang 2007), and is the topic of this paper.

Proper scoring rules were introduced independently by Brier (1950), Good (1952, p. 112), and de Finetti (1962). They have traditionally been 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, many deviations from expected value maximization have been observed, and this may explain why proper scoring rules have not been used by modern decision theorists so far. The first deviation was pointed out by Bernoulli (1738), who noted 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 besides utility curvature (Kahneman & Tversky 1979; 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"). Then phenomena occur that are fundamentally different than those for known probabilities, which adds to the descriptive failure of expected value. Gilboa (1987), Gilboa & Schmeidler (1989), Hogarth & Einhorn (1990), Schmeidler (1989), and Tversky & Kahneman (1992) developed decision theories that incorporate ambiguity. Halevy (2007) provided a recent empirical study into the new phenomena. Typical new implications for economic theory are in Hansen, Sargent, & Tallarini (1999) and Mukerji & Tallon (2001).

It is high time that proper scoring rules be updated 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, where violations of expected value have been widely documented. Such an update, provided by this paper, brings mutual benefits for practitioners of proper scoring rules and for the study of risk and ambiguity. For practitioners of proper scoring rules we show how to improve the empirical performance and validity of their measurement instrument. For studies of risk and ambiguity we show how to benefit from the efficient measurement instrument provided by proper scoring rules. Regarding the first benefit, we bring bad news when describing the many empirical deviations from expected value that distort classical proper scoring rules, but good news when we give quantitative assessments of those distortions and ways to correct for them. In the experiment in this paper we will find no systematic biases for a repeated-payment treatment, so that no correction may be needed for group averages then. This can be further good news for classical applications of proper scoring rules. Regarding the second benefit, we show how subjective beliefs and ambiguity attitudes can easily be isolated from risk attitude, using the incentive compatibility and efficiency of proper scoring rules.

Our correction technique can be interpreted as a new calibration method (Keren 1991; Yates 1990) that does not need many repeated observations, unlike traditional calibration methods (Clemen & Lichtendahl 2005). An efficient aspect of our method is that we need not elicit the entire risk attitudes of agents so as to correct for them. For instance, we need not go through an entire measurement of the utility and probability weighting functions. Instead, we can immediately infer the correction from a limited set of readily observable data (the "correction curve"; see later).

We emphasize that the biases that we correct for need not concern mistakes on the part of our subjects. Deviations from risk neutrality need not be irrational and, according to some, even deviations from Bayesian beliefs need not be irrational, nor are the corresponding ambiguity attitudes (Gilboa & Schmeidler 1989). Thus, learning and incentives need not generate the required corrections. The required corrections concern empirical deficiencies of the model of expected value, i.e. they concern the researchers analyzing the data.

We illustrate the feasibility of our method through 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 confirm the usefulness of our method. Violations of additivity of subjective beliefs are reduced but not eliminated by our corrections. Thus, the classical measurements will contain violations of additivity that are partly due to the incorrect assumption of expected value, but partly they are genuine. Subjective beliefs are genuinely nonadditivity. They cannot be modeled through additive subjective probabilities.

From the Bayesian perspective, violations of additivity are undesirable. Because we can measure these violations, we can investigate which of several implementations of proper scoring rules best approximate the Bayesian model. To illustrate this point, we compared two experimental treatments: (1) only one single large decision is randomly selected and paid for real; (2) every decision is paid, and subjects earn the sum of (moderate) payments. Because of the law of large numbers one expects the results of treatment (2), with repeated small payments, to stay closer to expected value and Bayesianism than those of treatment (1) will. This was confirmed in our experiment, where smaller corrections were required for the repeated payments than for the single payment.

The analysis of this paper consists of three parts. The first part (§§3-5) considers various modern theories of risk and ambiguity, and derives implications for proper scoring rules from these theories. This part is of interest to practitioners of proper scoring rules because it shows what distortions affect these rules. It is of interest to decision theorists because it shows a new field of application.

The second part of the paper, §§6-7, applies the revealed-preference reversal technique to the results of the first part. That is, we do not assume theoretical models to derive empirical predictions therefrom, but we assume empirical observations and derive the theoretical models from those. §6 presents the main result of this paper, showing how subjective beliefs can be derived from observed choices in an easy manner. §7 contains a simple example illustrating such a derivation at the individual level. 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 only in applying our method empirically can skip most of §§3-6, reading only §3 up to Theorem 3.1 and

Corollary 6.4. For practitioners of proper scoring rules, the second part of this paper then shows how beliefs can be derived from observed proper scoring rules under more realistic descriptive theories. We introduce so-called risk-corrections to correct for distortions. For the study of subjective (possibly non-Bayesian) beliefs and ambiguity attitudes, the second part of this paper shows how proper scoring rules can be used to measure and analyze these concepts efficiently. An observed choice in a proper scoring rule gives as much information as an observed indifference in a binary choice while avoiding the empirical difficulties of observing indifferences.

The third part of the paper, §§8-11, presents an experiment where we implement our correction method. We present some preliminary findings on nonadditive beliefs and on different implementations of real incentives. For brevity, detailed examinations of empirical implementations of our method, of the descriptive and normative properness of additive subjective beliefs, the effects of real incentives, and also of interpretations of beliefs and ambiguity attitudes, are left as topics for future study. Our contribution is to show how those concepts can be measured. The experiment of this paper, thus, serves to demonstrate the empirical implementability of our theoretical contribution.

§8 contains methodological details. §9 presents results regarding the biases that we correct for, and §10 presents some implications of the corrections of such biases. Discussions and conclusions are in §§11-12. 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 100th digit of π is 3, and so on. 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 an event E. Mostly these have been derived from: (a) binary preferences, which only give inequalities or approximations; (b) binary indifferences, which are hard to elicit, e.g. through the complex Becker-DeGroot-Marschak mechanism (Braga & Starmer 2005; Karni & Safra 1987) or bisection (Abdellaoui, Vossman, & Weber (2005); (c) introspection, which is not revealed-preference based let alone incentive-compatible. 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.

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

$$(E: 1-(1-r)^2, 1-r^2), \qquad (2.1)$$

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). The following symmetry between E and E^c will be crucial in later theories.

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

$$r_{E^c} = 1 - r_E. \qquad (2.2)$$

3. Proper Scoring Rules and Subjective Expected Value

The first two parts of our analysis concern a theoretical analysis of proper scoring rules. This section considers the model commonly assumed for proper scoring rules, from their introduction in the 1950s up to today: *subjective expected value* maximization. It means, first, that the agent assigns a subjective probability p to each event E.[2] Second, the agent maximizes expected value with respect to probabilities.

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

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

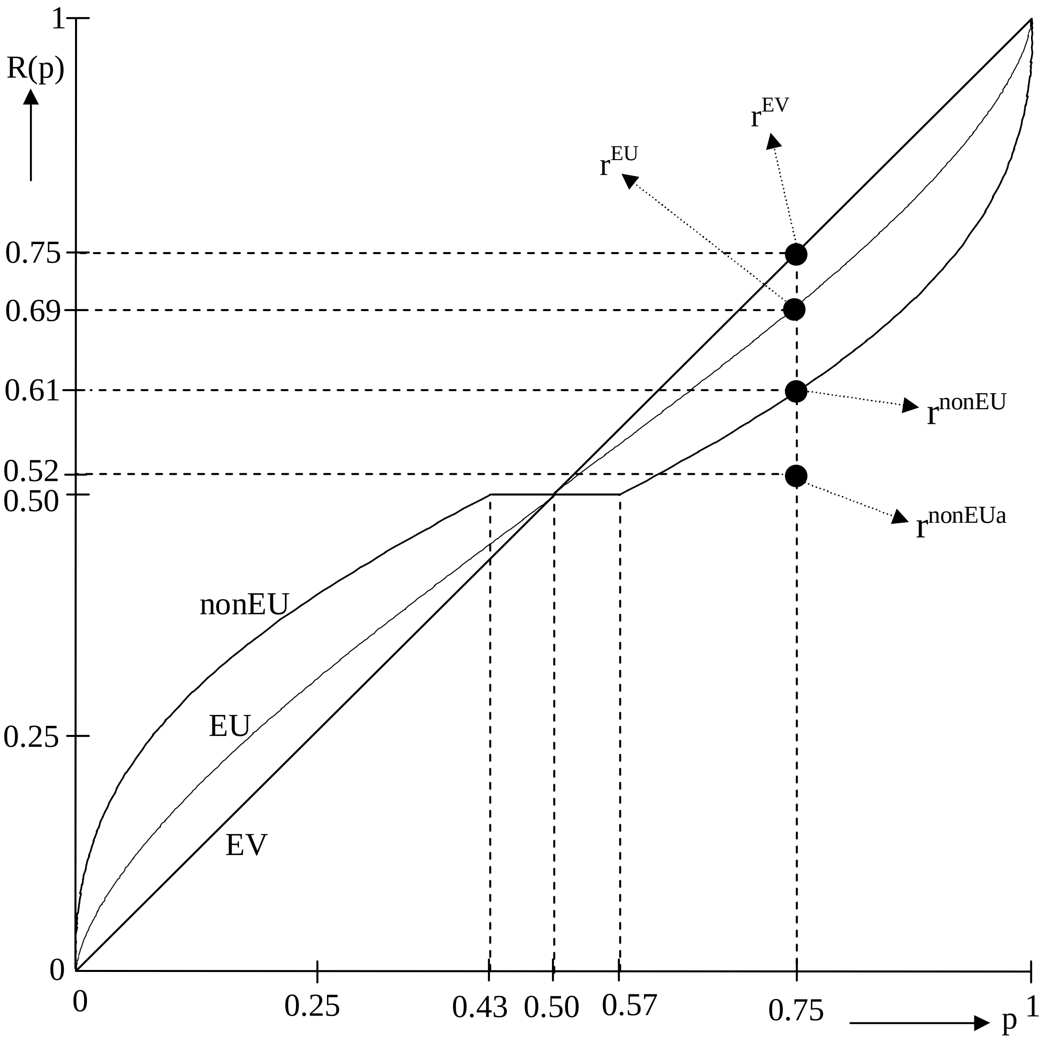
If event E has probability p, then we also write R(p) for r_E throughout this paper. According to Eq. 3.1, and all other models considered in this paper, all events E with the same probability p have the same value r_E, so that R(p) is well-defined. We have the following corollary of Eq. 2.2.

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

[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 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.

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 discussed in what follows.

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 4.3.4.
r^{nonEUa}: nonexpected utility for unknown probabilities ("Ambiguity").

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

It is in the agent's best interest to truthfully report his subjective probability of E. This explains the term “reported probability.” In Theorem 3.1, reported probabilities satisfy the Bayesian additivity condition for probabilities. *Additivity* is the well-known property that the probability of a disjoint union is the sum of the separate probabilities. We call the number r_E the (*uncorrected*) *reported probability*.

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 two parts of this paper, we use variations of the following theoretical example.

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. Two Commonly Found Deviations from Expected Value under Risk, and Their Implications for Quadratic Proper Scoring Rules

This section considers two factors that distort proper scoring rule measurements, and that should be corrected for. These factors concern decision attitudes and can be identified from decision under risk, with events for which probabilities are given. Proper scoring rules serve to examine other kinds of events, namely events with unknown probabilities. Those events will be the topic of the following sections. This section considers known probabilities only so as to identify biases, as a preparation for the following sections. §4.1 defines the domain of decision under risk, and then explains the organization of the other subsections.

4.1. Decision under Risk

ASSUMPTION 4.1.1. [Decision under Risk]. For event E, an objective probability p is given. □

Any deviation of a reported probability r_E from the objective probability p entails a bias that should be corrected for. Under expected value maximization we obtain, similarly as in Theorem 3.1, that the agent should report r=p, so that there is no bias. The hypothetical situation of an agent using a subjective probability in Theorem 3.1 different than the objective probability in Assumption 4.1.1 cannot arise under plausible assumptions.[3] Subjective probabilities agree with objective probabilities whenever the latter exist, and this will be assumed throughout.

The effects of the factors that deviate from expected value and that distort the classical proper scoring rule measurements, explained later, are illustrated in Figure 3.1. Their quantitative size will be illustrated through extensions of Example 3.2. §4.2 considers the first factor generating deviations, being nonlinear utility under expected utility. This section extends earlier studies of this factor by Winkler & Murphy (1970). We use expected utility and its primitives as in Savage's (1954) usual model. Extensions to alternative models (Broome 1990; Karni 2007; Luce 2000) are a topic for future research. §4.3 considers the second factor, namely violations of expected utility for known probabilities.

4.2. The First Deviation: Utility Curvature

Bernoulli (1738) put forward the first deviation from expected value. Because of the risk aversion in the so-called St. Petersburg paradox, Bernoulli proposed that people maximize the expectation of a *utility function* U. We assume that U is continuously differentiable with

[3] The first assumption is what defines decision under risk: that the only relevant aspect of events is their objective probability, and the second that we have sufficient richness of events to carry out the following reasoning. The claim then 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).

positive derivative everywhere, implying strict increasingness. We assume throughout that U(0) = 0. Eq. 3.1 is now generalized to

$$pU(1-(1-r)^2) + (1-p)U(1-r^2) . \tag{4.2.1}$$

The first-order optimality condition for r, and a rearrangement of terms (as in the proof of Theorem 5.2), implies the following result. For r ≠ 0.5, the theorem also follows as a corollary of Theorem 5.2 and Eq. 3.2.

THEOREM 4.2.1. Under expected utility with p the probability of event E, the optimal choice r = R(p) satisfies:

$$r = \frac{p}{p + (1-p)\frac{U'(1-r^2)}{U'(1-(1-r)^2)}} . \tag{4.2.2}$$

□

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); it is confirmed empirically by Huck & Weizsäcker (2002). The decision-based distortion in the direction of 0.5 is opposite to the overconfidence (probability judgments too far from 0.5) mostly found in direct judgments of probability without real incentives (McClelland & Bolger 1994), and found among experts seeking to distinguish themselves (Keren 1991, p. 2f24 and 252; the “expert bias”, Clemen & Rolle 2001). Optimistic and pessimistic distortions of probability can also result from nonlinear utility if the probability considered is a consensus probability for a group of individuals with heterogeneous beliefs (Jouini & Napp 2007).

EXAMPLE 4.2.2. Consider Example 3.2, but assume expected utility with $U(x) = x^{0.5}$. Substitution of Eq. 4.2.2 (or Theorem 5.2 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 a risk premium, designating a profit margin for an insurance company. By Eq. 2.2, $r_C = 0.31$, and by symmetry $r_G = r_S = r_Y = 0.31$ too. The reported probabilities violate additivity, because r_G +

$r_S + r_Y = 0.93 > 0.69 = r_E$. This violation in the data reveals that expected value does not hold. □

OBSERVATION 4.2.3. Under expected utility with probability measure P, $r_E = 0.5$ implies P(E) = 0.5. Conversely, P(E) = 0.5 implies $r_E = 0.5$ if risk aversion holds. Under risk seeking, $r_E \neq 0.5$ is possible if P(E) = 0.5. □

Theorem 4.2.1 clarifies the distortions generated by nonlinear utility, but it does not provide an explicit expression of R(p), i.e. r as a function of p, or vice versa. It seems to be impossible, in general, to obtain an explicit expression of R(p). We can, however, obtain an explicit expression of the inverse of R(p), i.e. p in terms of r (Corollary 6.1). For numerical purposes, R(p) can then be obtained as the inverse of that function—this is what we did in our numerical analyses, and how we drew Figure 3.1.

4.3. 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.

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

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

We now turn to the second deviation from expected value. 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 (Machina 1987; Starmer 2000; Sugden 2004). For the prospects relevant to this paper, QSRs with only two outcomes and no losses, all presently popular static 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 $p \geq 0.5$.

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

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.

$$\text{For } r < 0.5\text{: } w(1-p)U(1-r^2) + (1 - w(1-p))U(1-(1-r)^2). \qquad (4.3.2)$$

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. 4.3.1 and in Eq. 4.3.2, 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 4.3.1. We display the implication.

OBSERVATION 4.3.2. 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]

THEOREM 4.3.3. Under nonexpected utility with p the probability of event E, the optimal choice r = R(p) satisfies:

$$\text{For } r > 0.5\text{:}\;\; r = \frac{w(p)}{w(p) + (1-w(p))\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}\;. \tag{4.3.3}$$

□

The above result, again, follows from the first-order optimality condition, and also follows as a corollary of Theorem 5.2 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 section (Eqs. 5.6 and 5.7).

EXAMPLE 4.3.4. Consider Example 4.2.2, but assume nonexpected utility with $U(x) = x^{0.5}$ and

$$w(p) = \Big(exp(-(-ln(p))^{\alpha})\Big) \tag{4.3.4}$$

with parameter $\alpha = 0.65$ (Prelec 1998). This function agrees with common empirical findings (Tversky & Kahneman 1992; Abdellaoui 2000; Bleichrodt & Pinto 2000; Gonzalez & Wu 1999). From Theorem 4.3.3 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 4.2.2 generated by w for this event E has led to an extra distortion of r_E by 0.08. The extra expected-value loss (and, hence, the extra risk premium) relative to Example 4.2.2 is 0.8094 – 0.7920 = 0.0174. By Eq. 4.3.1, $r_C = 0.39$, and by symmetry $r_G = r_S = r_Y = 0.39$ too. The reported probabilities strongly violate additivity, because $r_G + r_S + r_Y = 1.17 > 0.61 = r_E$. □

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

Figure 3.1 illustrates the effects through the curve indicated by nonEU. 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 4.2.3). 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.

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

5. A Third Commonly Found Deviation from Subjective Expected Value Resulting from Non-Bayesian Beliefs and Ambiguity, and Its Implications for Quadratic Proper Scoring Rules

This section considers a third deviation from expected value maximization. This deviation does not (merely) concern decision attitudes as did the two deviations examined in the preceding section. It rather concerns subjective beliefs about events with unknown probabilities (which involves ambiguity). These are the events that proper scoring rules serve to examine. Thus, the deviation in this section does not concern something we necessarily have to correct for, but rather it concerns something that we want to measure and investigate without a commitment as to what it should look like.

In applications of proper scoring rules it is commonly assumed that the agent chooses (Bayesian) subjective probabilities p = P(E) for such events, where these subjective probabilities are assumed to satisfy the laws of probability. The agent evaluates prospects the same way for subjective probabilities as if these probabilities were objective, leading to the following modification of Eq. 4.3.1:

$$\text{For } r \geq 0.5\text{: } w(P(E))U(1-(1-r)^2) + (1-w(P(E)))U(1-r^2). \qquad (5.1)$$

For w the identity with w(P(E)) = P(E), Eq. 5.1 reduces to subjective expected utility, the subjective version of Eq. 4.2.1. In applications of proper scoring rules it is commonly assumed that not only w, but also U is the identity, leading to subjective expected value maximization, the model analyzed in §3.

The approach to unknown probabilities of Eq. 5.1, treating uncertainty as much as possible in the same way as risk, is called *probabilistic sophistication* (Machina & Schmeidler 1992). All results of §4 can be applied to this case, with distortions generated by nonlinear U and w. Probabilistic sophistication can be interpreted as a last attempt to maintain Bayesianism at least at the level of beliefs. Empirical findings, initiated by Ellsberg (1961), have demonstrated however that probabilistic sophistication is commonly violated empirically.

EXAMPLE 5.1 [Violation of Probabilistic Sophistication]. Consider Example 4.3.4, 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 4.3.4. Under probabilistic sophistication combined with nonexpected utility as in Example 4.3.4, r_{E_a} must be the same as r_E in Example 4.3.4 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 fails. This observation also suggests that r_{E_a} may differ from r_E. □

[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.

The deviations from expected value revealed by Ellsberg through 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 components of beliefs and decision attitudes that are typical of unknown probabilities. They 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).

As explained in Appendix B, virtually all presently existing models for decision under uncertainty evaluate the qsr-prospect of Eq. 2.1 in the following way:

$$\text{For } r \geq 0.5\text{: } W(E)U(1-(1-r)^2) + \big(1-W(E)\big)U(1-r^2). \tag{5.2}$$

Here W is a nonadditive set function often called *weighting function* or capacity, which satisfies the natural requirements that W assigns value 0 to the vacuous event $\varnothing$, value 1 to the universal event, and is increasing in the sense that $C \supset D$ implies $W(C) \geq W(D)$. For completeness, we also give the formula for $r < 0.5$, which can be obtained from Eq. 5.2 through symmetry (Observation 2.1).

$$\text{For } r < 0.5\text{: } \big(1-W(E^c)\big)U(1-(1-r)^2) + W(E^c)U(1-r^2). \tag{5.3}$$

Under probabilistic sophistication (Eq. 5.1), subjective belief P can be recovered from W through $P = w^{-1}(W)$, where w^{-1} can be interpreted as a correction for non-neutral risk attitudes. Machina (2004) argued that almost-objective probabilities can be constructed in virtually all circumstances of uncertainty, so that a domain for w is always available. In general, the “risk-corrected” function $w^{-1}(W)$ need not be a probability. We write

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

In general, B is what remains if the risk component w is taken out from W. It is common in decision theory to interpret factors beyond risk attitude as ambiguity. Then B reflects ambiguity attitude. There is no consensus about the extent to which ambiguity reflects non-Bayesian beliefs, and to what extent it reflects non-Bayesian decision attitudes beyond belief. If the equality $B(E) + B(E^c) = 1$ (*binary additivity*) is violated, then it can further be debated whether B(E) or $1 - B(E^c)$ is to be taken as an index of belief or of ambiguity. Such interpretations have not yet been settled, and further studies are called for. We will usually refer to B as reflecting beliefs, to stay as close as possible to the terminology used in the literature on proper scoring rules today. On some occasions we will refer to the decision-theoretic ambiguity. Irrespective of the interpretation of B, it is clear that the behavioral component w^{-1} of risk attitude should be filtered out before an interpretation of belief can be considered. This paper shows how this filtering out can be done.

In Schmeidler (1989), the main paper to initiate Eqs. 5.2 and 5.3, w was assumed linear, with expected utility for given probabilities, and W coincided with B. Schmeidler interpreted this component as reflecting beliefs. So did the first paper on nonadditive measures for decision making, Shackle (1949). Many studies of direct judgments of belief have supported the thesis that subjective beliefs may deviate from Bayesian probabilities (McClelland & Bolger 1994; Shafer 1976; Tversky & Koehler 1994). Bounded rationality is an extra reason to expect violations of additivity at the level of beliefs (Aragones et al. 2005; Charness & Levin 2005).

We rewrite Eq. 5.2 as

$$\text{For } r \geq 0.5\text{: } w(B(E))U(1-(1-r)^2) + \big(1-w(B(E))\big)U(1-r^2). \qquad (5.4).$$

$$\text{For } r < 0.5\text{: } \big(1-w(B(E^c))\big)U(1-(1-r)^2) + w(B(E^c))U(1-r^2). \qquad (5.5)$$

In general, B assigns value 0 to the vacuous event ∅, value 1 to the universal event, and B is increasing in the sense that $C \supset D$ implies $B(C) \geq B(D)$. These properties similarly hold for the composition $w(B(\cdot))$, as we saw above.

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 4.3.1. 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

$$(E: 2-(1-r)^2, 1-r^2), \qquad (5.6)$$

i.e. qsr-prospects as in Eq. 2.1 but with a unit payment added under event E. The classical proper-scoring-rule properties of §2 are not affected by this modification, and the results of §3 are easily adapted. With this modification, we have the liberty to combine event E with the highest outcome both if E is more likely than E^c and if E is less likely, and we avoid the restriction of Observation 4.3.1. We then can observe w of the preceding subsection, and W(E) and B(E) over their entire domain. Similarly, with prospects

$$(E: 1-(1-r)^2, 2-r^2), \tag{5.7}$$

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. We reveal their biases according to the current state of the art of decision theory, suggest remedies whenever possible, and signal the problems that remain. Further investigations of the, we think promising, modifications of QSRs in the above equations are left 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 events less likely than their complements are 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 so far.

THEOREM 5.2. Under Eq. 5.4, the optimal choice r satisfies:

$$\text{If } r > 0.5\text{, then } r = r_E = \frac{w(B(E))}{w(B(E)) + \big(1-w(B(E))\big)\dfrac{U'(1-r^2)}{U'(1-(1-r)^2)}}\ . \tag{5.8}$$

□

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. 5.8 will be relatively small, i.e. close to 0.5. We give a numerical example.

EXAMPLE 5.3. Consider Example 5.1. 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 5.2, r_{E_a} will be smaller than r_E. Given the strong aversion to unknown probabilities that is often found empirically (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 4.3.4 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 4.3.4 is 0.7920 – 0.7596 = 0.0324. This amount can be interpreted as the ambiguity-premium component of the total uncertainty premium. By Eq. 4.3.1, $r_C = 0.48$, 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 5.2 is valid for virtually all static models of decision under uncertainty and ambiguity known in the literature today, because Eqs. 5.4 and 5.5 capture virtually 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 $W = w \circ B$.

6. Measuring Beliefs through Risk Corrections

The next two sections, constituting the second part of the analysis of this paper, analyze proper scoring rules using the revealed-preference technique. That is, we do not derive empirical predictions from theoretical models, but we reverse the implication. We assume that empirical observations are given and derive theoretical models from these. In particular, we will derive beliefs B(E) from reported probabilities r_E. Before turning to this technique, we discuss alternative measurements of beliefs B considered in the literature.

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

$$B(E) = w^{-1}(W(E)). \tag{6.1}$$

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) or elaborate nonparametric measurements (Abdellaoui, Vossman, & Weber 2005).

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

$$(p:x, y) \sim (E:x, y) \tag{6.2}$$

for some preset $x > y$, say $x = 100$ and $y = 0$. Then $w(B(E))\big(U(x)–U(y)\big) = w(p)\big(U(x)–U(y)\big)$, and $B(E) = p$ follows. Wakker (2004) discussed the interpretation of Eqs. 6.1 and 6.2 as belief. Canonical probabilities were commonly used in early decision analysis (Raiffa 1968, §5.3; Yates 1990 pp. 25-27) under the assumption of expected utility. A recent experimental measurement is in Holt (2006, Ch. 30), who also assumed expected utility. Abdellaoui, Vossman, & Weber (2005) measured and analyzed them in terms of prospect theory, as does our paper. A practical difficulty is that the measurement of canonical probabilities requires the measurement of indifferences, and these are not easily inferred from choice. For example, Holt (2006) used the Becker-deGroot-Marschak mechanism, and Abdellaoui, Vossman, & Weber (2005) used a bisection method. Huck & Weizsäcker (2002) compared the QSR to the measurement of canonical probabilities and found that the former is more accurate.

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

We now use the revealed-preference technique to introduce risk corrections. These combine the advantages of measuring $B(E) = w^{-1}(W(E))$, of measuring canonical probabilities, and of calibrating reported probabilities relative to objective probabilities, while avoiding the problems described above, by benefiting from the efficiency of proper scoring rules. The QSR does entail a restriction of the observations regarding B(E) to cases of E being more likely than E^c (Observation 4.3.1). The first results do express beliefs B (or p) in terms of observed values r, but are not complete revealed-preference results because the right-hand sides of the equations still contain utilities, which are theoretical quantities that are not directly observable. A "coincidental" agreement of two right-hand sides along the way will then lead to the main result of this paper: a complete revealed-preference result, deriving beliefs B entirely from observable choice.

We first consider expected utility of §4.2. The following result follows from Theorem 4.2.1 through algebraic manipulations or, for $r \neq 0.5$, as a corollary of Corollary 6.2 hereafter.

COROLLARY 6.1. Under expected utility with p the (objective or subjective) probability of event E, and $r = R(p)$ the optimal choice, we have

$$p = \frac{r}{r + (1-r)\frac{U'(1-(1-r)^2)}{U'(1-r^2)}} \ . \tag{6.3}$$

□

We next consider nonexpected utility for known probabilities as in §4.3. An explicit expression of p in terms of (U and) r, i.e. of $R^{-1}(p)$, follows next for $r > 0.5$, assuming that we can invert the probability weighting function w. The result follows from Theorem 4.3.3.

COROLLARY 6.2. Under nonexpected utility with given probabilities (Eq. 4.3.1), the optimal choice $r = R(p)$ satisfies:

$$\text{If } r > 0.5, \text{ then } p = R^{-1}(r) \ = \ w^{-1}\left(\frac{r}{r + (1-r)\frac{U'(1-(1-r)^2)}{U'(1-r^2)}}\right) . \tag{6.4}$$

□

In general, it may not be possible to derive both w and U from R(p) without further assumptions, i.e. U and w may be nonidentifiable for proper scoring rules. Under regular assumptions about U and w, however, they have some different implications. The main difference is that, if we assume that U is differentiable (as done throughout this paper) and concave, then a flat part of R(p) around 0.5 must be caused by w (Observation 4.2.3).

We, finally, turn to nonexpected utility if no probabilities are known, as in §5. Theorem 5.2 implies the following results. 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 6.3. Under nonexpected utility with unknown probabilities (Eq. 5.4), the optimal choice $r = r_E$ satisfies:

$$\text{If } r > 0.5, \text{ then } B(E) \ = \ w^{-1}\left(\frac{r}{r + (1-r)\frac{U'(1-(1-r)^2)}{U'(1-r^2)}}\right). \qquad (6.5)$$

□

As a preparation for a complete revealed-preference result, note that the right-hand sides of Eqs. 6.4 and 6.5 are identical. Hence, if we find a p in Eq. 6.4 with the same r value as E, then we can, by Eq. 6.4, immediately substitute p for the right-hand side of Eq. 6.5, getting B(E) = p without need to know the ingredients w and U of Eq. 6.5. This observation (to be combined with Eq. 2.2 for r < 0.5) implies the following corollary, which is displayed for its empirical importance and which is the main result of this paper.

COROLLARY 6.4. Under nonexpected utility with unknown probabilities (Eq. 5.4), assume for the optimal choice $r = r_E$ that $r > 0.5$. Then

$$B(E) = R^{-1}(r). \qquad (6.6)$$

□

This corollary is useful for empirical purposes. It is the only implication of our theoretical analysis that is needed for applications. It shows how proper scoring rules can allow for deviations from expected value and expected utility, and is key in filtering out risk attitudes. 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. 6.6. 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.

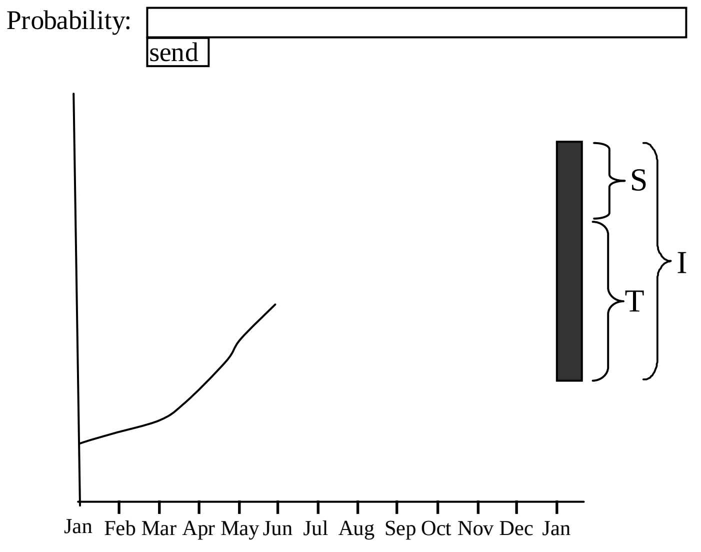
We call the function R^{-1} the *risk-correction* (for proper scoring rules), and $R^{-1}(r_E)$ the *risk-corrected probability*. This value is the canonical probability, obtained without having measured indifferences such as through the Becker-DeGroot-Marschak mechanism, without having measured U and 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. We can then conclude with more confidence that overconfidence is a genuine property of belief, irrespective of risk seeking.

7. 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. We will see that Corollary 6.4 is the only result of the theoretical analysis needed to apply our method. Results and curves for r < 0.5 are derived from r > 0.5 using Eq. 2.2; we will not mention this point explicitly in what follows.

FIGURE 7.1. Layout of the screens



The left side of Figure 7.1 displays the performance of stock 12 in our experiment from January 1 until June 1 1991 as given to the participants. Stock 12 concerned the Begemann Kon. Groep (General Industries). Further details (such as the absence of a unit on the y-axis) will be explained in §8. 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 14, the results are as follows.

$$r_S = 0.35;\ r_T = 0.55;\ r_I = 0.65. \tag{7.1}$$

Under additivity of reported probability, $r_S + r_T - r_I$ (the *additivity bias*, defined in general in Eq. 8.5), should be 0, but here it is not and additivity is violated.

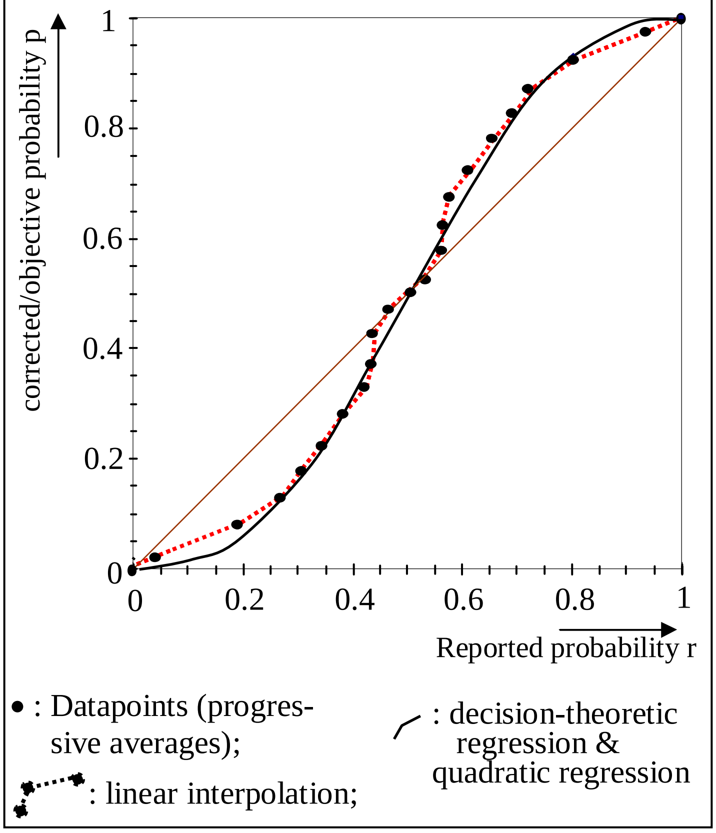
$$\text{The additivity bias is } 0.35 + 0.55 - 0.65 = 0.25. \tag{7.2}$$

Table 7.1 and Figure 7.2 (in inverted form) display the reported probabilities R(p) that we measured from this participant, with the curves explained later. We use progressive averages (midpoints between data points) so as to reduce noise.[6]

TABLE 7.1. Progressive average reported probabilities R(p) of participant 14

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

FIGURE 7.2. $R^{-1}(r)$ of participant 14



For simplicity of presentation, we analyze the data here using linear interpolation. Then R(0.23) = 0.35.[7] Using this value for R(0.23), using the values R(0.56) = 0.55 and R(0.77) = 0.65, and, finally, using Eq. 6.6, we obtain the following risk-corrected beliefs.

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

$$B(S) = R^{-1}(0.35) = 0.23;\ B(T) = R^{-1}(0.55) = 0.56;\ B(I) = R^{-1}(0.65) = 0.77;$$
$$\text{the additivity bias is } 0.23 + 0.56 - 0.77 = 0.02. \qquad (7.3)$$

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

It is statistically preferable to fit data with smoother curves than resulting from linear interpolation. We derived “decision-theoretic” parametric curves for R(p) from Corollary 6.2, with further assumptions explained at the end of §9.1.[8] The resulting curve for participant 14 is given in the figure. The equality $B = R^{-1}(r)$ and this curve lead to

$$B(S) = R^{-1}(0.35) = 0.24;\ B(T) = R^{-1}(0.55) = 0.59;\ B(I) = R^{-1}(0.65) = 0.76;\ \text{the additivity bias is } 0.24 + 0.59 - 0.76 = 0.07, \qquad (7.4)$$

again reducing the uncorrected additivity bias. For this participant the quadratic curve, explained in §11, happens to be indistinguishable from the decision theoretic curve.

8. An Experimental Application of Risk Corrections: Method

The following four sections present the third part of this paper, being an experimental implementation of our new measurement method.

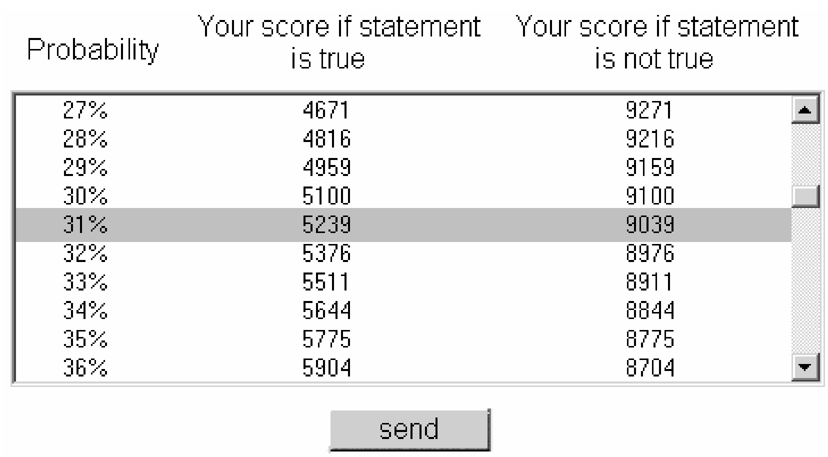
Participants. N = 93 students from a wide range of disciplines (45 economics; 13 psychology, 35 other disciplines) participated in the experiment. They were self-selected from a mailing list of approximately 1100 people.

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

[8] The decision-theoretic curve in the figure is the function $p = B(E) = \dfrac{r}{r + (1-r)\dfrac{0.26\big(1-(1-r)^2\big)^{-1.26}}{0.26(1-r^2)^{-1.26}}}$, in agreement with Corollaries 6.2 and 6.4, where we estimated $w(p) = p$ and found $\rho = -0.26$ as optimal value for U(x) in Eq. 8.1.

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 8.1.

FIGURE 8.1.



Probability	Your score if statement is true	Your score if statement is not true
27%	4671	9271
28%	4816	9216
29%	4959	9159
30%	5100	9100
31%	5239	9039
32%	5376	8976
33%	5511	8911
34%	5644	8844
35%	5775	8775
36%	5904	8704

All figures (including Figure 7.1) 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 the "evaluation date." The uncertain events always concerned the question whether or not the price of a stock would lie in a target-interval seven months after the evaluation date. For each stock, the participants received a graph depicting the price of the stock on 0, 1, 2, 3, 4, and 5 months prior to the evaluation date, as well as an upper and lower bound to the price of the stock on the evaluation date. Figure 7.1, 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 7.1). 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 large interval for each 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 8.1, 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 4.3.2. 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. 4.3.1. Notice that this equation can be estimated using a non-parametric procedure. If the agent is willing to go through a large series of correction questions, it is possible to measure the corresponding reported probability of each objective probability repeatedly. In this way an accurate estimate of the whole correction curve can be obtained without making assumptions about the utility function or the weighting function. This procedure seems the appropriate one if the goal is to correct an expert, e.g., correct the reports provided by a weatherman. In applications of experimental economics where subjects participate for a limited amount of time, the researcher will only be able to collect a limited number of observations of the correction curve. Then it is more appropriate to follow a parametric approach to elicit the curve that fits the observations best. In this paper, we used parametric fittings. For U we used the *power utility with parameter* ρ, also known as the family of constant relative risk aversion (CRRA)[9], and the most popular parametric family for fitting utility, which is defined as follows:

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

For $\rho>0$: $U(x) = x^\rho$;

for $\rho=0$: $U(x) = ln(x)$;

for $\rho<0$: $U(x) = -x^\rho$. (8.1)

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

$$w(p) = \big(exp(-\beta(-ln(p))^\alpha)\big), \tag{8.2}$$

chosen for its analytic tractability and good empirical performance. We will mostly use the one-parameter subfamily with $\beta=1$, as in Eq. 4.3.4, for reasons explained later. Substituting the above functions yields

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

for Eq. 6.5.

The model we estimate is as follows.

$$R_{s,t,k}(j/20) = h(j/20,\alpha_t, \rho_t) + \varepsilon_{s,t,k}(j/20,\sigma_t^2)\,. \tag{8.3}$$

Here $R_{s,t,k}(j/20)$ is the reported probability of participant s for known probability $p=j/20$ ($10 \le j \le 19$) in treatment t (t = ALL or t = ONE) for the kth measurement for this probability, with only k=1 for j = 10, k = 1,2 for $11 \le k \le 18$, and k = 1,2,3 for j = 19. With β set equal to 1, α_t is the remaining probability-weighting parameter (Eq. 8.2), and ρ_t is the power of utility (Eq. 8.1). The function h is the inverse of Eq. 6.4. Although we have no analytic expression for this inverse, we could calculate it numerically in the analyses. The error terms $\varepsilon_{s,t,k}(j/20)$ are drawn from a truncated normal distribution with mean 0 and treatment dependent variance σ_t^2. The distribution of the error terms is truncated because reported probabilities below 0 and above 1 are excluded by design. Error terms are identically and independently distributed across participants and choices. We employed maximum likelihood to estimate the parameters of Eq. 8.3. We also carried out an analysis at the individual level of the calibration part, with $\alpha_{s,t}$ and $\rho_{s,t}$ instead of α_t and ρ_t, i.e. with these parameters depending on the participant.

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

$$r_S + r_T = r_I. \tag{8.4}$$

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

$$B(S) + B(T) \ = \ B(I). \tag{8.5}$$

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 average additivity bias, as relevant for aggregated group behavior and expert opinions, was enlarged or reduced by correction; (c) whether the absolute value of the additivity bias, as relevant for additivity at the individual level, was enlarged or reduced by correction. We report only the tests for averages over all stocks.

9. Results of the Calibration Part

Risk-corrections and, in general, QSR measurements, do not make sense for participants who are hardly responsive to probabilities, so that R(p) is almost flat on its entire domain. Hence we kept only those participants for whom the correlation between reported probability and objective probability exceeded 0.2. We thus dropped 4 participants. The following analyses are based on the remaining 89 participants.

9.1. Group Averages

We did several tests using Eq. 8.2 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 9.1 lists the estimates for the model of Eq. 8.3 for β=1 (Eq. 4.3.4 instead of Eq. 8.2) together with the estimates of some models with additional restrictions. We first give results for group averages, assuming homogeneous participants.

TABLE 9.1. 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.

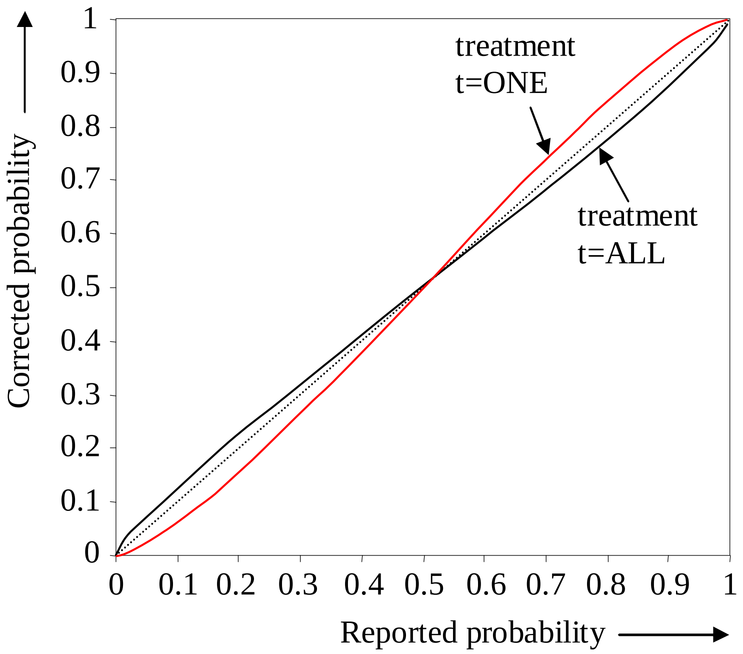
Overall need for risk-correction. The 1st row of Table 9.1 shows the results for the most general model. The 2nd row presents the results without any correction. The likelihood reduces significantly (Likelihood Ratio test, $p = 0.01$) and substantially, so that risk-correction is called for. Risk-correction 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 9.1, 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, $\rho_{ONE} = \rho_{ALL} = 1$).

FIGURE 9.1. Corrected versus reported probability



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 parameter 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.

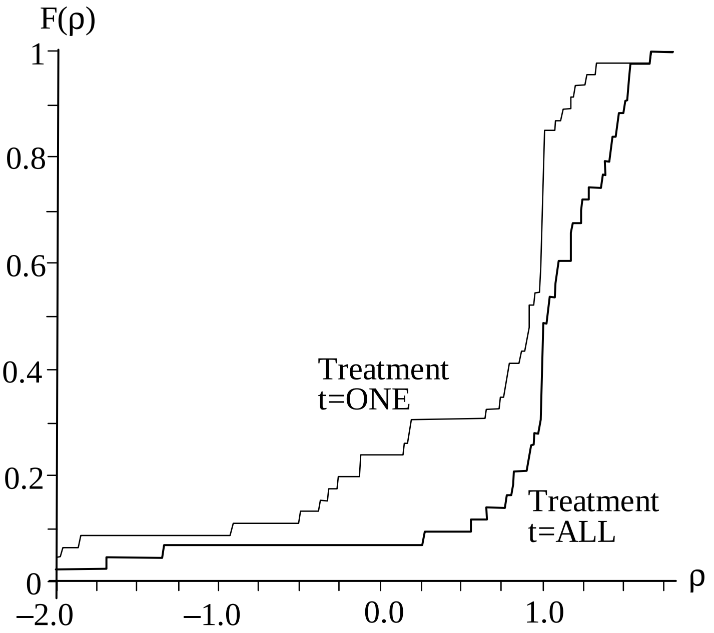
We found that ρ has a slightly better explanatory power than α. For this reason, and for reasons of convenience (see discussion section), we will only use the parameter ρ, and assume α = 1 henceforth. Figure 9.1 displays the resulting average risk-correction for the two treatments separately.

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 9.1 to its 9th row ($\rho_{ALL} = 1$) (Likelihood Ratio test, p = 0.01).

9.2. 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 9.2, 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 ρ=1 (i.e., no correction) on both sides. As seen from the group-average analysis, there are more deviations at the risk-averse side of ρ < 1.

FIGURE 9.2. Cumulative density of ρ

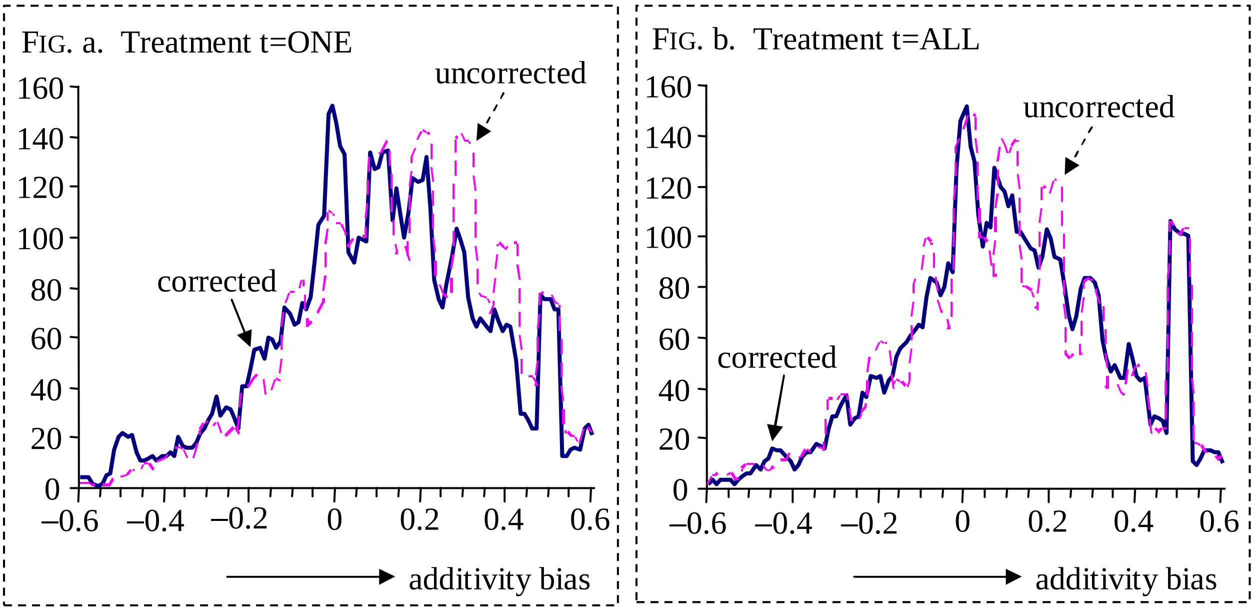


Comparing the two treatments. The ρ-coefficient distribution of treatment t=ONE dominates the ρ-coefficient distribution of treatment t=ALL. A two-sided Mann-Whitney test rejects the null-hypothesis that the ranks of ρ-coefficients are equal across the treatments in favor of the hypothesis that the ρ-coefficients 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).

10. Results for the Stock-Price Part: Risk-Correction and Additivity

FIGURE 10.1. 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 10.1 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, in agreement with common findings in the literature (Tversky & Koehler 1994; Bateman et al. 1997). Indeed, for virtually all stocks the additivity bias is significantly positive for both treatments, showing in particular that additivity does not hold. This also holds when taking the average additivity bias over all stocks as one data point per participant ($z = 5.27$, $p < 0.001$ for t=ONE, $z = 4.35$, $p < 0.001$ for t=ALL). We next consider whether risk corrections reduce the violations of additivity.

Let us first consider 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 10.1 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 the average additivity bias over all stocks for each individual, with no significant differences generated by correction (average 0.128 uncorrected and average 0.136 corrected; $z = -1.64$, $p = 0.1$). Similar results hold for absolute values of additivity biases, which gave a significant increase for 1 stock and a significant decrease for no stock. Taking the average additivity bias over all stocks for each individual (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.

11. Discussion of Experiment

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

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

Optimal Incentive Scheme. 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; Harrison et al. 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 that is commonly generated by repeated choice.[10] For the purpose of measuring individual preference, the treatment t=ONE is, therefore, preferable. When the purpose is, however, to derive subjective probabilities from proper scoring rules, and no risk-correction is possible, then a drift towards expected value is actually an advantage, because uncorrected proper scoring rules assume expected value. This point agrees with our findings, where less risk-correction was required for the t=ALL treatment. Li (2007) discusses other arguments for and against repeated rewarding when events are not verifiable and when binary rewards have to be used.

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

Choice of Parameters. The lack of extra explanatory power of parameter β in Eq. 8.2 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 4.3.2).

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

Pragmatic applications. More tractable families can be used to fit the reported probabilities than the decision-theory-based curves that we used. For example, in Figure 7.2 we also used quadratic regression to find the curve $p = a + br + cr^2$ that best fits the data. For most participants, the curve is virtually indistinguishable from the decision-theoretic curve. This observation, together with Corollary 6.4 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 analysis of this paper, and the decision-theory based curve-fitting that we adopted, served to prove that our method is in agreement with modern decision theories. If this thesis is accepted, and the only goal is to obtain risk-corrected reported probabilities, then one may choose the pragmatic shortcuts just described.

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.

The experimental data show that for a subset of the subjects a substantial correction of reported probabilities needs to be made. The fraction of the population that needs substantial corrections is larger when only one big decision is paid than when repeated small decisions are paid. Our conclusion is that it is desirable to correct agents' reported probabilities elicited with scoring rules, especially if only a single large decision is paid. If it is not possible to obtain individual measurements of the correction curve, then it will be useful to use best-guess corrections, for instance through averages obtained from individuals as similar as possible. Thus, at least, the systematic error for the group average to risk attitude has been corrected for as good as is possible without requiring extra measurements. In this respect the average curves in our Figure 9.1 are reassuring for existing studies, because these curves suggest that only small corrections were called for regarding the group averages in our context.

Allen (1987) proposed to avoid biases of the QSR resulting from nonlinear utility by paying in terms of the probability of winning a prize instead of in terms of money, and this procedure was implemented by McKelvey & 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 enhance the desired risk neutral behavior.

12. Conclusion

This paper has applied modern theories of risk and ambiguity to proper scoring rules. Mutual benefits have resulted for practitioners of proper scoring rules and for the study of risk and ambiguity. For the former we have shown which distortions affect their common measurements and how large these distortions are, using theories that are descriptively better than the expected value hypothesis still common for proper scoring rules today. We have provided a procedure to correct for the aforementioned distortions, and a theoretical foundation has been given for interpretations of the resulting measurements as (possibly non-Bayesian) beliefs and/or ambiguity attitudes. For studies of risk and ambiguity we have shown how the remarkable efficiency of proper scoring rules can be used to measure and analyze subjective beliefs and ambiguity attitudes in ways more tractable than is possible through the binary preferences traditionally used.

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

Appendix A. Proofs and Technical Remarks

In Eqs. 4.3.1 and 4.3.2, 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. 5.4 and 5.5, 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 4.2.3. If $r = 0.5$ then the marginal utility ratio in Eq. 4.2.2 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. 4.2.2 would be at least 1 so that the right-hand side of Eq. 4.2.2 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 5.2. We write π for the decision weight W(E). For optimality of interior solutions r, the first-order optimality condition for Eq. 5.4 is that

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

implying

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

or $\pi U'(a-b(1-r)^2) = r \times \big(\pi U'(a-b(1-r)^2) + (1-\pi)U'(a-br^2)\big)$, and Eq. 5.8 follows.

□

PROOF OF COROLLARY 6.3. Let $r > 0.5$ be optimal, and write $\pi = W(E)$. Then Eq. A.1 implies

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

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

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 6.3 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 6.4. Theorem 5.2 implies that the right-hand side of Eq. 5.8 is r both as is, and with p substituted for B(E). Because Eq. 5.8 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. 5.4 and 5.5 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), and Wakker & Deneffe (1996, p. 1143 and pp.1144-1145).

The convenient feature that most decision theories agree on the evaluation of binary prospects was pointed out by Miyamoto (1988), calling Eqs. 5.4 and 5.5 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, Ch. 3) and in Gonzalez & Wu (2003). Only for more than two outcomes, the theories diverge (Mosteller & Nogee 1951 p. 398; Luce 2000, introductions to Chs. 3 and 5). Theories that also deviate for two outcomes include betweenness models (Chew & Tan 2005), the variational model (Maccheroni, Marinacci, & Rustichini (2006), and models with underlying multistage decompositions (Halevy &

Feltkamp 2005; Halevy & Ozdenoren 2007; Klibanoff, Marinacchi, & Mukerji 2005; Nau 2006; Olszewski 2007).

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

This appendix will appear on internet after publication, and is not meant to be incorporated in the publication.

[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 then 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.

Allais, Maurice (1953), “Le Comportement de l’Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l’Ecole Américaine,” *Econometrica* 21, 503–546.

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

Aragones, Enriqueta, Itzhak Gilboa, Andrew Postlewaite, & David Schmeidler (2005), “Fact-Free Learning,” *American Economic Review* 95, 1355–1368.

Bateman, Ian J., Alistair Munro, Bruce Rhodes, Chris Starmer, & Robert Sugden (1997), “Does Part-Whole Bias Exist? An Experimental Investigation,” *Economic Journal* 107, 322–332.

Bernoulli, Daniel (1738), “Specimen Theoriae Novae de Mensura Sortis,” *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 5, 175–192.

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

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.

Broome, John R. (1990), “Bolker-Jeffrey Expected Utility Theory and Axiomatic Utilitarianism,” *Review of Economic Studies* 57, 477–502.

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

Charness, Gary & Dan Levin (2005), "When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect," *American Economic Review* 95, 1300–1309.

Chew, Soo Hong & Guofu Tan (2005), "The Market for Sweepstakes," *Review of Economic Studies* 72, 1009–1029.

Clemen, Robert T. & Kenneth C. Lichtendahl (2005), “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, 3.

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.

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.

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.

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,” mimeo.

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

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

Halevy, Yoram (2007), “Ellsberg Revisited: An Experimental Study,” *Econometrica*, forthcoming.

Halevy, Yoram & Vincent Feltkamp (2005), “A Bayesian Approach to Uncertainty Aversion,” *Review of Economic Studies* 72, 449–466.

Halevy, Yoram & Emre Ozdenoren (2007), “Uncertainty and Compound Lotteries: Calibration,” working paper, University of British Columbia.

Hansen, Lars Peter, Thomas J. Sargent, & Thomas D. Tallarini (1999), "Robust Permanent Income and Pricing," *Review of Economic Studies* 66, 873–908.

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

Harrison, Glenn W., Morten I. Lau,& M.B. Williams (2002),"Estimating Individual Discount Rates in Denmark: A Field Experiment," *American Economic Review* 92, 1606–1617.

Hogarth, Robin M. & Hillel J. Einhorn (1990), "Venture Theory: A Model of Decision Weights," *Management Science* 36, 780–803.

Hogarth Robin M. & Howard C. Kunreuther (1985), "Ambiguity and Insurance Decisions," *American Economic Review, Papers and Proceedings* 75, 386–390.

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

Holt, Charles A. (2006),"*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.

Huck, Steffen & Georg Weizsäcker (2002), "Do Players Correctly Estimate What Others Do? Evidence of Conservatism in Beliefs," *Journal of Economic Behavior and Organization* 47, 71–85.

Johnstone, David J. (2006), "The Value of Probability Forecast from Portfolio Theory," School of Business, University of Sydney, Australia.

Jouini, Elyès & Clotilde Napp (2007), "Consensus Consumer and Intertemporal Asset Pricing with Heterogeneous Beliefs," *Review of Economic Studies* 74, 1149–1174.

Kahneman, Daniel & Amos Tversky (1979), "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* 47, 263–291.

Karni, Edi (2007), "A New Approach to Modeling Decision-Making under Uncertainty, *Economic Theory* 33, 225–242.

Karni, Edi & Zvi Safra (1987), "Preference Reversal and the Observability of Preferences by Experimental Methods," *Econometrica* 55, 675–685.

Karni, Edi & Zvi Safra (1989), "Dynamic Consistency, Revelations in Auctions and the Structure of Preferences," *Review of Economic Studies* 56, 421–434.

Keren, Gideon B. (1991), "Calibration and Probability Judgments: Conceptual and Methodological Issues," *Acta Psychologica* 77, 217–273.

Keynes, John Maynard (1921), "*A Treatise on Probability.*" McMillan, London.

Klibanoff, Peter, Massimo Marinacci, & Sujoy Mukerji (2005), "A Smooth Model of Decision Making under Ambiguity," *Econometrica* 73, 1849–1892.

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

Li, Wei (2007), "Changing One's Mind when the Facts Change: Incentives of Experts and the Design of Reporting Protocols," *Review of Economic Studies* 74,

Luce, R. Duncan (1991), "Rank- and-Sign Dependent Linear Utility Models for Binary Gambles," *Journal of Economic Theory* 53, 75–100.

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

Maccheroni, Fabio, M. Marinacci, & A Rustichini (2006), "Ambiguity Aversion, Robustness, and the Variational Representation of Preferences," *Econometrica* 74, 1447–1498.

Machina, Mark J. (1987), "Choice under Uncertainty: Problems Solved and Unsolved," *Journal of Economic Perspectives* 1 no 1, 121–154.

Machina, Mark J. (2004), "Almost-Objective Uncertainty," *Economic Theory* 24, 1–54.

Machina, Mark J. & David Schmeidler (1992), "A More Robust Definition of Subjective Probability," *Econometrica* 60, 745–780.

Manski, Charles F. (2004), "Measuring Expectations," *Econometrica* 72, 1329–1376.

McClelland, Alastair & Fergus Bolger (1994), "The Calibration of Subjective Probabilities: Theories and Models 1980–1994." *In* George Wright & Peter Ayton (eds.), *Subjective Probability*, 453–481, Wiley, New York.

McKelvey, Richard & Talbot Page (1986), "Common Knowledge, Consensus, and Aggregate Information," *Econometrica* 54, 109–127.

Miyamoto, J.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.

Mukerji, Sujoy & Jean-Marc Tallon (2001), "Ambiguity Aversion and Incompleteness of Financial Markets," *Review of Economic Studies* 68, 883–904.

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.

Olszewski, Wojciech (2007), “Preferences over Sets of Lotteries,” *Review of Economic Studies* 74, 567–595.

Nau, Robert F. (2006), “Uncertainty Aversion with Second-Order Utilities and Probabilities,” *Management Science* 52, 136–145.

Palfrey, Thomas R. & Stephanie W. Wang (2007), “On Eliciting Beliefs in Strategic Games,” Division of the Humanities and Social Sciences, CalTech, Pasadena, CA 91125.

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

Prelec, Drazen (1998), “The Probability Weighting Function,” *Econometrica* 66, 497–527.

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. (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.

Schoemaker, Paul J.H. (1982), “The Expected Utility Model: Its Variations, Purposes, Evidence and Limitations,” *Journal of Economic Literature* 20, 529–563.

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.

Shackle, George L.S. (1949), “A Non-Additive Measure of Uncertainty,” *Review of Economic Studies* 17, 70–74.

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

Shiller, Robert J., Fumiko Kon-Ya, & Yoshiro Tsutsui (1996), “Why Did the Nikkei Crash? Expanding the Scope of Expectations Data Collection,” *The Review of Economics and Statistics* 78, 156–164.

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 (2000), "Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk," *Journal of Economic Literature* 38, 332–382.

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

Sugden, Robert (1991), "Rational Choice: A Survey of Contributions from Economics and Philosophy," *Economic Journal* 101, 751–785.

Sugden, Robert (2004), "Alternatives to Expected Utility." *In* Salvador Barberà, Peter J. Hammond, & Christian Seidl, *Handbook of Utility Theory, Vol. II*, 685–755, Kluwer Academic Publishers, Dordrecht.

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

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," *Psychological 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. & 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.