Network for Studies on Pensions, Aging and Retirement

Netspar DISCUSSION PAPERS

Stefan Trautmann and Gijs van de Kuilen

Belief Elicitation

A Horse Race Among Truth Serums

DP 11/2011-118

Belief Elicitation:
A Horse Race among Truth Serums

Stefan T. Trautmann & Gijs van de Kuilen*
Tilburg University

November 2011

In survey studies, probabilistic expectations about uncertain events are typically elicited by asking respondents for their introspective beliefs. If more complex procedures are feasible, beliefs can be elicited by incentive compatible revealed preference mechanisms ("truth serums"). Various mechanisms have been proposed in the literature, which differ in the degree to which they account for respondents' deviations from expected value maximization. In this paper, we pit non-incentivized introspection against five truth serums, to elicit beliefs in a simple two-player game. We test the internal validity (additivity and predictive power for own behavior), and the external validity (predictive power for other players' behavior, or accuracy) of each method. We find no differences among the truth serums. Beliefs from incentivized methods are better predictors of subjects' own behavior compared to introspection. However, introspection performs equally well as the truth serums in terms of accuracy and additivity.

KEY WORDS: belief measurement, subjective probability, scoring rules, outcome matching, probability matching, internal validity, external validity

JEL-CLASSIFICATION: D81, C83, C91

---

* Dept. of Economics, & CentER,, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, the Netherlands. s.t.trautmann@uvt.nl, G.v.d.Kuilen@uvt.nl. Dirk Engelmann and Peter Wakker gave helpful comments. Financial support by NWO-VENI grants to Stefan Trautmann and Gijs van de Kuilen is gratefully acknowledged.

1. Introduction

In decisions under uncertainty, information about the probabilities of the various events is often unavailable, and decisions have to be made on the basis of subjective probability judgments. Agents should form these subjective assessments of likelihoods (beliefs) according to the laws of probability, and evaluate alternatives by their belief-weighted expected utility (Savage 1954). Because of their importance in economic decisions, economists have elicited subjective beliefs in a wide range of applications. Game theorists have tested whether subjective beliefs about other players' behavior can explain deviations from Nash equilibrium strategies (e.g. Bellemare, Kroeger & van Soest 2008; Blanco et al. 2011; Costa-Gomez and Weizsäcker 2008; Rey-Biel 2009). Macroeconomists have studied the effect of beliefs about uncertain future income and demand on savings and investment decisions (Guiso, Japelli and Terlizzese 1992; Guiso and Parigi 1999). In development economics, researchers tried to link beliefs to decisions to adopt new variety of seed or to settle in natural disaster prone areas (Cameron and Shah 2011; Delavande, Gine, and McKenzie 2011), and health economists investigated whether wrong beliefs can explain risky health behaviors such as smoking or shunning of preventive care (Carman and Kooreman 2010; Ahmed Khwaja, Frank Sloan and Martin Salm, 2006; Khwaja et al. 2007). Although in theorizing it is typically assumed that agents form beliefs rationally, in empirical applications they may differ from the true unknown probabilities. If they differ, it has been shown that subjective beliefs are often better predictors of behavior than objective likelihoods, even if the latter are available to the researcher.[1]

Various methods have been employed to elicit beliefs. The most common method involves directly asking respondents about their introspective beliefs. Simple introspective questions can easily be included in surveys and are easy to explain to respondents. On the downside, respondents may have little incentive to think carefully about the problem, thus adding noise. More seriously, respondents may misrepresent their beliefs because of social desirability, or to influence what they believe is the goal of the research (Li 2007; Manski 2004; Zizzo 2010). To overcome the problems of simple introspection, methods have been developed that extract beliefs from respondents' revealed preferences between prospects that

[1] E.g. Branch (2004) for macroeconomic expectations; Armantier & Treich 2009; Bellemare, Kroeger & van Soest 2008; Costa-Gomez & Weizsaecker 2008; Heinemann, Nagel & Ockenfels 2009; Nyarko & Schotter 2002, for behavior in games; Haruvy, Lahav, & Noussair 2007 for trading in asset markets; Carman & Kooreman 2010 for medical decisions; Guiso, Japelli & Terlizzese 1992 for precautionary saving. See also Attanasio (2009) and Hurd (2009).

offer real monetary payoffs depending on the uncertain event of interest to the researcher. If the monetary incentives dominate other motives, these mechanisms are *incentive compatible*, i.e., reporting true beliefs maximizes the respondent's belief-weighted expected payoff (Hurwicz 1960). Because it is in the best interest of the respondent to reveal her beliefs truthfully, incentive compatible methods have sometimes been called *truth serums* (Prelec 2004). Many truth serums assume that respondents are risk-neutral expected utility maximizers. Because both risk neutrality and expected utility maximization may empirically be violated (Machina 1987; Starmer 2000), refinements have been developed that account for risk aversion and for non-linear probability weighting (Andersen et al. 2010; Heinemann et al. 2009; Hossain and Okui 2011; Offerman et al. 2009).

Given the importance of measuring people's subjective expectations in economics and other social sciences, an assessment of the costs and benefits of the different methods is warranted. While the cost of implementing a method in a specific setting (survey, experiment) can usually easily be assessed by the individual researcher, its relative benefits are less obvious. There is little evidence yet on whether the various truth serums improve the quality of the elicited beliefs over and above the benchmark of simple introspection, and if so, by what degree. Similarly, there is little evidence on whether more complex (and thus theoretically more robust) methods yield more reliable data. In many applications it may be impossible to condition monetary incentives on an event of interest, because it falls outside the period in which the researcher has access to the subjects. Even in laboratory experiments, the researcher may want to avoid complex belief elicitation tasks because they distract subjects form the main task of interest (Cabrales et al. 2010; Haruvy, Lahav, and Noussair 2007). In these cases it is important to understand potential deviations from true beliefs caused by introspective questions or less robust truth serums.

The aim of the current paper is to study in a controlled environment whether (1) truth serums provide higher quality belief data than simple introspection, and (2) whether more complex truth serums are worth the additional effort because they elicit better data than less sophisticated truth serums. In contrast to the previous literature, we consider a large set of truth serums, and compare the methods using three different quality criteria. Applications may be concerned with different aspects of the subjective beliefs, and differences in the rank-order of the methods across the three criteria provide information on which methods best serve the goal of the researcher. To assess the benefits of different truth serums, the current paper pits introspection against five incentive compatible mechanisms of different degree of

complexity.[2] In a between-subject design, we compare introspection to three widely used truth serums, viz. the *outcome matching* method (e.g., Kadane and Winkler 1988; Heinemann et al. 2009; Huck and Weizsäcker 2002), the *probability matching* method (e.g., Abdellaoui, Vossmann and Weber 2005; Arrow 1951, Hollard, Massoni, and Vergnaud 2010; Holt 2006), and the *quadratic scoring rule* (e.g., Brier 1950; Costa Gomes and Weizsäcker 2008; McKelvey and Page 1990; Nyarko and Schotter 2002; Rey-Biel 2009). Moreover, we consider a correction method for outcome matching that controls for deviations from risk neutrality under expected utility used by Heinemann et al. (2009), and a correction method for the scoring rule that controls for deviations from risk neutrality, caused by either utility curvature or probability weighting introduced by Offerman et al. (2009). We elicit players' beliefs in a simple two-person ultimatum game, and measure the internal and external validity of each method. *External validity* concerns in how far the elicited beliefs match the true objective probabilities of the event (accuracy). As a measure of external validity we thus study the predictive power of the elicited beliefs for the behavior of the other players in the game. *Internal validity* concerns "the degree to which persons give internally consistent, sensible responses to the questions" (Manski 2004, p. 1343). As measures of internal validity we employ the additivity of the elicited probabilities (Attanasio 2009; Tversky and Koehler 1994), and the consistency of players' beliefs with their own behavior in the game (Costa-Gomez and Weizsäcker 2008; Costa-Gomez et al. 2008).

A few studies have directly compared subsets of the widely used methods, employing as a quality criterion the external validity (accuracy) of the elicited beliefs. Hoa and Houser (2010) compare two implementations of probability matching proposed by Karni (2009). Friedman and Massaro (1998), Sonnemans and Offerman (2001), and Rutstrom and Wilcox (2009), compare non-incentivized introspection with the quadratic scoring rule. Hollard, Massoni, and Vergnaud (2010) include introspection, the quadratic scoring rule and probability matching in their study. Huck and Weizsäcker (2002) compare the quadratic scoring rule and outcome matching. The evidence from this literature is mixed, and shows no clear advantage of incentivized methods over introspection. Similarly, no truth serum seems to dominate the others in terms of accuracy. While Huck and Weizsäcker find that the scoring rule performs better than outcome matching, in Hollard et al.'s study the scoring rule gives worse results than either introspection or probability matching. However, none of these

[2] Complexity refers the implementation of the task (number of questions asked, payment mechanism, and difficulty of instructions, and randomizations).

studies considered corrections for risk attitudes in the scoring rule or outcome matching methods. Consequently, deviations from risk neutrality have been brought forward to explain poor performance of these truth serums in some studies. Hossain and Okui (2011) propose a new scoring rule method that is insensitive to deviations from expected value maximization, and show that it elicits more accurate beliefs than the quadratic scoring rule. Andersen et al. (2010) compare the quadratic and the linear scoring rule, and elicit a large number of risky choices alongside the belief questions. They find no differences between the two methods, but show that correcting for risk attitudes influences the estimated subjective beliefs. These studies support the view that risk corrections are important in belief elicitation.

Building on the existing literature, the current paper studies a larger set of methods, including corrections for outcome matching and the scoring rule. The current paper also extends previous analyses by using criteria of both external and internal validity to compare truth serums and introspection. The existing evidence is completely silent on the comparative performance of methods in terms internal validity measures.[3] We compare the *empirical performance* of the different methods, being agnostic about the *theoretical validity* of each method in our current setting and subject pool. We believe that in virtually all empirical applications, the researcher has little information about validity of the various assumptions underlying the methods (precise risk attitudes, violations of expected utility), and is interested only in eliciting high quality data. For example, if a method that does not control for incentive compatibility leads to smaller violations of additivity than a method that does control for incentive compatibility, we consider the former, theoretically less robust, method empirically preferable.

The current paper finds little evidence for improved empirical performance of more complex methods. There are no robust differences among the five truth serums for any of the three criteria considered. Moreover, the results show that introspective belief measurement performs similar in terms of accuracy and additivity as the truth serums. However, incentivized beliefs are a better predictor of players' own behavior in the game. The current study thus supports the view that introspection is a valid method to measure subjective beliefs in applications where accuracy and additivity are important. If more complex methods are

[3] Armantier and Treich (2009) employ introspection, a scoring rule and a prediction contest in their study of beliefs first price auctions. They show that subjective probabilities explain overbidding, and report that the elicitation method did not affect their results. Their study thus provides indirect evidence for the equivalence of introspection and scoring rule beliefs terms of internal validity.

feasible, there can be benefits from using incentive compatible mechanisms to predict agents' behavior.

The paper proceeds as follows. Section 2 gives definitions and notation. Section 3 introduces the belief elicitation methods that are considered in the current study, followed by the experimental design in Section 4. Experimental results are presented in Section 5, and discussed in Section 6. Section 7 concludes.

2. Definitions and Notation

We study beliefs in the context of the ultimatum game. In stage 1 of the game, a *proposer* proposes the division of an amount of €20 between herself and another player, called the *responder*. The proposer chooses from a menu of six possible allocations of the type (proposer receives, responder receives), viz. (€20, €0), (€16, €4), (€12, €8), (€8, €12), (€4, €16), and (€0, €20). In stage 2, the responder decides whether to accept or to reject the proposal. If she accepts, the proposal is implemented. If she rejects, both players receive €0.

We elicit full strategies for responders in the ultimatum game. That is, we let responders indicate for each possible proposal whether she accepts it or rejects it. The game is resolved by matching the actual proposal to the respective strategy of the responder. Before resolving the game, we elicit proposers' beliefs about the rejection/acceptance probabilities for each of the six proposals, and elicit responders' beliefs about the proposal.

Let E denote an uncertain event, such as whether the responder accepts the offer (€20, €0). We call an event *risky* if objective probabilities are exogenously given, i.e., when the probability p of event E is known a priori. In many situations no objective probabilities are available to the decision maker. We call an event *uncertain* in this case. In the ultimatum game, events based on behavior of the other player are uncertain.

In discussions of the theoretical properties of the truth serums, we assume that under uncertainty agents form subjective assessments of likelihoods of the uncertain events (beliefs), according to the laws of probability. That is, we assume that agents are probabilistically sophisticated (Machina and Schmeidler 1992). Moreover, agents evaluate alternatives by the belief-weighted expected utility that each alternative yields (Savage 1954). More specifically, let $x_E y$ denote 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, and x and y designating monetary amounts in Euro. Under expected utility, the prospect is evaluated by $B(E)U(x) + B(E^c)U(y)$, where B(E) denotes the subjective belief in event E, U is a continuous and strictly

increasing utility function, and $B(E^c) = 1 - B(E)$. We call the prospect x_E0 a *bet* on event E, paying *x* if E obtains and zero otherwise. The *certainty equivalent* (CE) of a prospect is the certain amount that makes the agent indifferent between receiving the prospect or receiving the certain amount. The *matching probability* p of event E is defined by the indifference between a prospect paying *x* if event E obtains (and nothing otherwise), and a prospect paying *x* with probability p (and nothing otherwise), i.e., $x_E0 \sim x_p0$.

3. Introspection, Truth Serums, and the Cost of Implementation

This study measures beliefs in a simple ultimatum game with a restricted choice set, eliciting full strategies. This allows us to observe incentivized beliefs for strategies that are rarely implemented in actual play. We employ four methods to elicit beliefs of proposers and responders in the ultimatum game in a between subject design (introspection plus three truth serums). After applying corrections for risk attitudes to two of the methods, we obtain in total 6 different measurements of beliefs, which we describe in section 3.1. In section 3.2 we illustrate the differences in the cost of implementing each method, which underly the need to provide empirical estimates of the benefits of each method, to make an informed tradeoff between costs and benefits.

3.1. Measuring Beliefs: The Elicitation Methods

Introspection. In the introspection task each proposer was asked to state her beliefs concerning the rejection behavior of the responder in an introspective way. For each of the six proposals, proposers stated the probability that the allocation would be rejected by the responder by reporting a number between 0 and 100. Similarly, responders were asked to report for each proposal the probability that it was chosen by the proposer. Subjects received a fixed payment of €5 for this task.[4]

Outcome Matching. With outcome matching, beliefs were inferred from indifference between a sure amount of money and the prospect x_Ey, i.e,

$$CE \sim x_Ey \qquad (1)$$

[4] Details of the payment procedures for all methods are given in the experimental design section.

for some amounts $x > y$ (e.g., Kadane and Winkler 1988, Heinemann et al. 2009). Under the assumption of expected value maximization it follows from the equality CE = B(E)x + (1−B(E))y that the belief B(E) is equal to (CE–y)/(x–y). In the experiment, for each of the six proposals, we presented proposers with 21 choices between a prospect that pays €15 if the proposal was accepted (event E) and zero otherwise, and some sure amount. The larger the subjective probability of acceptance, the more attractive the prospect becomes. The sure amount varied between €0 and €15 in equally sized steps, and the 21 choices were presented in a choice list (see appendix). For low sure amounts the prospect is commonly chosen, while for large sure amounts the sure amount is commonly preferred. The midpoint between the values for which the subject switched between a preference for the prospect to a preference for the sure amount is taken as the certainty equivalent from which the subjective belief can be calculated. Similarly, for responders, for each proposal, a prospect paying €15 if the proposal was chosen by the proposer and zero otherwise was compared to 21 sure amounts in a choice list as described above. Again, the larger the subjective probability that the proposer accepts the proposal, the more attractive the prospect becomes, and the larger the elicited certainty equivalent should be.

Probability matching. In the probability matching task, we elicited for each subject the probability p that satisfies the indifference

$$x_p y \sim x_E y \qquad (2)$$

for some x > y. Under a general expectation function with w(p) transforming (subjective) probabilities and U(x) transforming outcomes, w(B(E))U(x) + (1-w(B(E)))U(y) = w(p)U(x)+ (1-w(p)U(y) implies B(E) = p.[5] Thus, probability matching is valid under a wide range of decision models, including expected utility, rank-dependent utility, and prospect theory. Probability matching was commonly used in early decision analysis (Raiffa 1968, §5.3; Yates 1990 pp. 25-27). More recent empirical measurements of beliefs through probability matching can be found in Holt (2006, Ch. 30), and in Abdellaoui, Vossman, and Weber (2005). In the experiment we incentivized probability matching by using choice lists for both players and for each allocation as discussed for outcome matching above (see appendix).

[5] From w(B(E))U(x)+(1-w(B(E)))U(y)=w(p)U(x)+(1-w(p)U(y) it follows that w(B(E))=w(p). The conclusion that B(E)=p thus hinges on the assumption that both subjective and objective probabilities are transformed equally (Wakker 2004).

Now, the choice lists involved choices between a prospect based on the decision of the other player (event E), and a prospect with a given probability p of winning the €15. There were again 21 choices in each list, with probability p ranging from 0 to 1 in equally-sized steps of .05. Probabilities were presented in terms of frequencies using a 20-sided die.

Quadratic scoring rule. Beliefs can be measured through so-called *scoring rules*, introduced by Brier (1950) and de Finetti (1962), which have the advantage that a single choice suffices to determine the exact belief in an incentive compatible way. Scoring rules have been used in many domains, including education (Echternacht 1972), finance (Shiller, Kon-Ya, and Tsutsui 1996), political science (Tetlock 2005), and experimental game theory (Costa Gomes and Weizsaecker 2008). The most popular scoring rule is the quadratic scoring rule (QSR; McKelvey and Page 1990, Nyarko and Schotter 2002). When rewarded according to the quadratic scoring rule, the agent is offered the prospect

$$[a - b(1 - r)^2]_{E}[a - br^2], \qquad (3)$$

with $a>0$ and $b > 0$. The parameter $r \in [0, 1]$ is chosen by the agent. A clairvoyant who knows that event E will obtain chooses $r = 1$ to maximize her payoff; a clairvoyant who knows that event E will not obtain chooses $r = 0$. A decision maker who is uncertain about the event E chooses $r = B(E)$ to maximize her expected payoff.[6]

In the experiment, beliefs were elicited with the quadratic scoring rule by presenting subjects with a table that calculates for each of 21 values of r, $r \in \{0, .05, ..., .95, 1\}$, the payoff under event E and under its complement E^c according to Eq.3 (see appendix). The parameters were set to a=€20 and b=€20, implying a symmetric payoff structure and a risk free payoff of €15 resulting from r=.5. Subjects chose one row in the table, rather than directly reporting their subjective probabilities, and received a payoff depending on which event materialized. Thus, for a proposer who believes that the proposal (€0; €20) is very likely to be accepted, it will be attractive to choose a row that yields a larger payoff under the event "proposal accepted" than under its complement. Note that subjects were not told that truthful reporting of their belief maximizes their expected payoff, or any related claim referring to truthful revelation being in their best interest. Although a similar list format was

[6] This follows immediately from the first order condition of expected value maximization with respect to r, $2bB(E)(1-r) = 2br(1-B(E))$.

used as in the outcome and probability matching methods, in the scoring rule task respondents made only a single decision.

Corrections for risk attitude. It is easy to see that in contrast to the probability matching method, the outcome matching method and the scoring rule are not incentive compatible if the respondent is not risk neutral. In the case of the scoring rule, for example, choosing r=.5 perfectly hedges the agent. If the agent is risk averse, the perfectly hedged prospect may be preferred over a risky one deriving from her true belief. Because risk attitude is typically found to deviate from risk neutrality, we consider refinements of these two methods that account for these deviations.

For the quadratic scoring rule we used the refinement proposed by Offerman et al. (2009) that corrects reported beliefs for risk aversion caused by utility curvature and non-linear probability weighting.[7] The method involves eliciting subjects' parameter r for risky events with known probability (e.g. a dice roll in our experiment). If the objective probability equals p, a deviation away from r = p indicates that the respondent hedges due to risk aversion.[8] Offerman et al. (2009, sec. 11.4.) show that beliefs can be corrected by fitting the non-linear function $p = \delta r+\gamma r^2$ (called the *correction function*) at the individual level using a set of objective probabilities p and the corresponding values of r reported by the subject. This yields estimated parameters δ^* and γ^*. Under uncertainty, the corrected beliefs then follow from the r_E, reported for the uncertain event E of interest, as $B(E) = \delta^* r_E+\gamma^*(r_E)^2$.

For outcome matching, under expected utility with utility function U(x) the elicited belief equals $B(E) = (U(CE) - U(y))/(U(x) - U(y))$ (e.g. Heinemann et al. 2009). Hence, for concave utility the ratio between numerator and denominator becomes larger compared to risk neutrality. Uncorrected beliefs would thus be biased downwards under risk aversion. We measured the utility function from a risky decision task and corrected reported beliefs for utility curvature in outcome matching.

3.2. Cost of Implementation

We have argued that more theoretical robustness comes at higher implementation costs, which suggests a tradeoff between cost and potential empirical benefits. In Table 1 we

[7] Kothiyal et al. (2011) generalize the correction techniques introduced by Offerman et al. (2009), and discuss methods to increase incentives for beliefs close to r=.5.

[8] Some deviations would obviously indicate risk seeking, but this is less common and poses no problems for the correction method.

summarize the implementation cost for the six measurement methods. Clearly, the truth serums require more effortful elicitation than simply asking subjects to report their belief. Moreover, theoretically more robust methods add additional costs in the form of multiple scoring rule tables, additional randomizations, or additional lottery choices. These implementation tools require more time and additional instructions, and potentially require more trust from the subjects if randomizations to play lotteries and to select choices/tables/choice lists are not easily observable by the subjects. The latter is often the case in computerized surveys and experiments.

TABLE 1: Implementing Truth Serums Experimentally

	Cost of implementation per event	Theoretical benefit: Valid under
Introspection	1 verbal question	-
Outcome Matching	1 choice list of 21 choices	EV
Probability Matching	1 choice list of 21 choices; 1 randomization with known probabilities	EV, EU, RDU, PT
Quadratic Scoring Rule	1 table of 1 choice	EV
Outcome Matching (corrected)	1 choice list of 21 choices; plus 1 risky choice question (once)	EV, EU
Quadratic Scoring Rule (corrected)	1 table of 1 choice; plus 10 tables of 1 choice each (once)	EV, EU, RDU, PT

Notes: EV=expected value; EU=expected utility maximizations; RDU=Rank-dependent utility; PT=prospect theory maximization

4. Experimental Design

General Procedures. Two-hundred-six undergraduate students from a wide range of disciplines participated in a computerized experiment that was conducted in the Tilburg University CentERlab.[9] The experiment had three stages. In stage 1, subjects played an ultimatum game using a full strategy method. The game was not resolved until the end of the experiment. Proposers' and responders' beliefs about the other player's behavior in the game were elicited in Stage 2. Stage 3 measured risk attitudes by a simple lottery choice task. For subjects in the scoring rule treatment, the third stage also elicited the parameters used for the

[9] The experiment used the z-Tree package by Fischbacher (2007).

correction method discussed in section 3. Subjects were given written instructions before each stage.[10]

No games, lotteries or belief-elicitation bets were resolved until the end of the experiment. After all decisions had been made by the subjects, one stage and one decision within this stage were randomly selected for each participant for real payment. This was done to prevent hedging and wealth effects (Starmer and Sugden 1991; Thaler and Johnson 1990; Blanco et al. 2010). The decisions were resolved and paid in private, and all randomizations for lotteries were done by throwing dice.

Stages and Treatments. The first stage was identical for all subjects. They were randomly assigned the role of either proposer or responder in the ultimatum game, and made their proposal or rejection/acceptance decisions as described in Section 2. The ultimatum game has been employed as a baseline task here because it provides a natural tension between the game theoretic prediction and fairness-related intuitions. All non-zero proposals should be accepted and thus the smallest non-zero amount been proposed for the responder if both players care only about monetary payoffs. Fairness considerations by either player will lead to deviations from this prediction. Because no player knows whether the other person considers such fairness issues in her choices, there is strategic uncertainty and beliefs become non-trivial. For proposers, the beliefs are an important input to their decision: given a belief that low proposals are rejected with high probability, even a proposer maximizing her expected monetary payoff (and thus not concerned with fairness) will offer a positive amount to the responder (Manski 2004, Bellemare et al. 2008).

The second stage belief measurement was implemented as a 4-treatment between-subject design. The treatments measured introspection (N=52), outcome matching (N=52), probability matching (N=50), and scoring rule (N=52) beliefs. Details of the different methods are discussed in section 3. For proposers we elicited beliefs about acceptance probabilities for each of the six proposals.[11] For responders, we elicited for each proposal the probability that it was chosen by the proposer. We additionally elicited for proposers their beliefs about the *rejection* probability for the proposal (€12, €8). We expected this proposal to yield subjective probabilities of both acceptance and rejection that were clearly bounded away from zero or one. For responders, we additionally elicited the probability that the

[10] Instructions and screen shots for different treatments can be found in the appendix.

[11] The six proposals of the type (proposer receives, responder receives) were (€20, €0), (€16, €4), (€12, €8), (€8, €12), (€4, €16), and (€0, €20). See Section 2 for details.

proposal (€12, €8) was *not* chosen by the proposer. These questions came after the questions regarding the acceptance (for proposers) and "was-chosen" (for responders) probabilities. For the proposal (€12, €8) we therefore obtained the subjective probability of the event E that the proposal was accepted/chosen, and the subjective probability of its complement E^c (proposal was rejected/not chosen). This allows us to study the additivity of the subjective beliefs, i.e. whether these belief measures add up to exactly 100%. Moreover, for responders we also study additivity of the choice probabilities over all six proposals.

The third stage of the experiment measured individual utility functions to control for deviations from risk neutrality; all subjects did the same task. The certainty equivalent of the prospect $10_{0.5}0$ was elicited by a simple choice list. For subjects in the scoring rule treatment, this stage contained an additional task. For these subjects we elicited scoring rule parameters r for 10 objectively known probabilities as described in section 3. This allows us to correct reported scoring rule beliefs for nonlinear risk attitude.

5. Results

This section presents the experimental results, and assesses the internal validity (do reported beliefs adhere to the laws of probability; do choices reflect subjective beliefs) and external validity (do reported beliefs successfully predict other players' behavior) of the elicitation methods used in the experiment. We first give the results of the ultimatum game and the utility estimation and risk attitude correction.

Ultimatum game. The probability that each allocation was chosen by the proposers, and the acceptance probabilities of responders for each allocation in the ultimatum game are given in Table 2. These probabilities serve as a benchmark for the analyses of the subjective beliefs below.

TABLE 2: Choice and Acceptance Probabilities Ultimatum Game

	Allocation					
	(€20, €0)	(€16, €4)	(€12, €8)	(€8, €12)	(€4, €16)	(€0, €20)
Choice probability	5.88%	19.61%	65.69%	6.86%	1.96%	0.00%
Acceptance probability	13.73%	43.14%	90.20%	95.10%	92.16%	88.24%

As typically found, proposers send positive amounts to responders with the modal proposal close to the equal split. On average, proposers send about 36% (€7.18) of the pie to responders, which is consistent with earlier findings (e.g. Roth 1995, chapter 4; Oosterbeek et al. 2004). Although positive offers in the ultimatum game can be interpreted in terms of social preferences (Fehr and Schmidt 2003), they can also derive from expected payoff maximization if the proposer believes that low offers are rejected (Manski 2004, Bellemare et al. 2008). Below we study if beliefs predict choices even in the absence of a social preference specification.

Responders' acceptance probabilities also showed a typical pattern by increasing from about 14% for the lowest offers to above 90% for proposals around the equal split. Acceptance was then decreasing slightly to 88.24% when the complete pie was offered to the responder. Thus, some responders accepted intermediate offers, but rejected low and high offers. This behavior has been found before, and can be explained by moral concerns of responders (e.g., Bellemare et al. 2008; Hennig-Schmidt et al. 2008).

Risk Attitudes and Corrections. In total, twelve subjects switched multiple times between the option yielding a certain amount of money and the option yielding the prospect $10_{0.5}0$. These observations were excluded from the analysis. The average certainty equivalent of the prospect $10_{0.5}0$ was €4.56. In total, 70.62% percent of subjects were risk averse, which is in line with what is commonly found in studies on individual decision making. For the parametric specification we assume the power utility function, $U(x) = x^{\rho}$, where $1-\rho$ is the coefficient of constant of relative risk aversion (CRRA), and $\rho=1$ indicates risk neutrality. The average (median) ρ in our sample is .978 (.931), which is significantly smaller than 1 (Wilcoxon signed-rank test, $z = 3.753$; $p<.01$). The elicited certainty equivalents in the outcome matching treatment were corrected for curvature of the utility function at the individual level using these CRRA estimates.

Reported beliefs in the scoring rule treatment were corrected for non-linear risk attitudes using the elicited correction function as discussed in section 3. Risk corrections do not make sense for respondents who are not responsive to objective probabilities, and we excluded 6 subjects for whom the correlation between objective probability and scoring rule parameter r was very low.[12] In Appendix C we give the results of the risk correction choices which

[12] Excluded subjects had correlations of zero (4 times), .13, and .20. The lowest correlation of included subjects was .71.

confirm that subjects' reported probabilities deviate from the objective known probabilities. The data also suggest that most deviations from the known probability were due to risk attitude and not pure noise. For both high and low probabilities, subjects are biased toward the .5 probability, but not beyond the .5 probability. For instance, for objective probability .8 we observed implied reported probabilities in the range .5 to .8, but virtually no such reported probabilities below .5. Thus, there seems to be little purely random choice, which would imply such a pattern (i.e., reported beliefs on the "other side" of the .5 probability).

Internal Validity of the Elicited Beliefs. As described in Section 4, for the proposal (€12, €8) we elicited both rejection and acceptance beliefs from proposers, and choice and not-choice beliefs from responders. Because the events are complements, their subjective probabilities should add up to 100 percent. This allows us to assess the internal validity of the different methods by testing whether elicited beliefs are additive. For responders we can also test whether the predicted choice probabilities over all 6 proposals add up to 100%.

TABLE 3: Additivity

	Introspection	Outcome Matching	Probability Matching	QSR	Outcome Matching (Corrected)	QSR (Corrected)
Proposers (€12, €8)	105 (105.5)*	120 (116)*	105 (113.8)*	110 (108.85)*	119.83 (118.22)**	109.94 (107.64)
Responders (€12, €8)	105 (106.65)*	105 (103)	100 (99.40)	102.5 (107.12)	106.65 (106.65)	103.61 (107.32)
Responders (6 proposals)	155.5 (183.27)**	185 (210)**	190 (199.80)**	227.5 (235.58)**	245.25 (226.29)**	203.67 (208.59)**

Notes: Entries medians (means) of the sum of the reported probabilities; */** significantly larger than 100% based on a two-sided Wilcoxon signed-ranks test, 5%/1% significance level.

The first row of Table 3 gives medians and means of the sum of proposers' reported probabilities of the event "(€12, €8) accepted" and its complement. There is a clear additivity bias, with the sum of the subjective probabilities significantly larger than 100% for all methods except the corrected scoring rule. Given the size of the deviations for the corrected scoring rule, however, this insignificance appears to be caused by larger variation rather than better additivity properties. The second row present results for the sum of the responders' reported probabilities of the event "(€12, €8) chosen" and its complement. Results are closer

to additivity, and deviations are insignificant for all truth serums.[13] The last row shows additivity of the responders' expected choice probabilities for all six proposals. We find that all deviations are large and significant, and that introspection performs significantly better than the quadratic scoring rule (Mann-Whitney-U test, $z=2.255$, $p<.05$) for the case of responders' beliefs over the 6 proposals. For this case, the correction of the quadratic scoring rule for deviations from risk neutrality improves additivity (Wilcoxon test, $z=3.412$, $p<.01$), replicating the finding in Offerman et al. (2009). We do not observe any other significant differences among the truth serums. The results show that elicited beliefs suffer from significant non-additivity, with the implied probabilities of the unions of events being too large. With 6 events considered, violations of additivity are more pronounced than for 2 events only. Overall there is little evidence of more complex methods outperforming simpler methods; in particular, introspection does not consistently underperform compared to the truth serums.

The second measure of internal validity concerns the consistency of proposers' beliefs with their own choices in the ultimatum game (Costa-Gomez and Weizsäcker 2008, Rey-Biel 2009). We calculate proposers' optimal choices in the game under the assumption that they maximize (1) expected value, (2) expected utility with CRRA utility function, or (3) a Fehr-Schmidt (1999) social preference utility function that considers differences in payoffs among the agents. The social preference function is parameterized using estimates in Bellemare et al. (2008) for all subjects. Results are in Table 4.

TABLE 4: Percentage of Proposers' Choices Consistent with Reported Beliefs

	Observed frequencies	Intro-spection	Outcome Matching	Probability Matching	QSR	Outcome Matching (Corrected)	QSR (Corrected)
EV	66**	19	32*	32*	31*	28	35*
EU	64**	17	28	30*	32*	32*	45**
FS	66**	19	28	36**	35*	24	35*

Note: Numbers are percentages. EV=Expected value maximization; EU= Expected utility maximization, FS=Fehr-Schmidt expected social preference function maximization; FS parameterized with $\alpha=0.85$ and $\beta=0.32$; EU is based on individual CRRA estimates. */** indicates significant better prediction than random at 5%/1% significance level.

Table 4 shows that the truth serums predict around 30-40% of the choices, which is significantly better than chance in most cases. Surprisingly, the performance is only mildly

[13] The same results obtain if we pool the data from proposers and responders.

affected by the utility function employed. Introspection performs poorly on this measure of consistency, with predictions correct in less than 20% of the cases. We do not find significant differences in performance among the truth serums. The table also includes predictions using the actual acceptance probabilities as shown in Table 2. We observe that in the current setting the objective probabilities (observed frequencies) predict behavior much better than the subjective beliefs. While objective probabilities are not available in many settings, the results show that subjective probabilities elicited by introspection may not improve predictions in comparison to simple random choice. The truth serums perform better, but still cannot account for much of the variation in actual behavior. The numbers are similar to the findings in Costa-Gomes and Weizsäcker (2008), while Rey-Biel (2009) finds higher rates of best-response to stated beliefs. Armantier and Treich (2009) find clear evidence that subjective probabilities are better predictors than objective probabilities in a first price auction experiment.

In sum, the internal validity of the subjective beliefs is affected little by the complexity of the elicitation method. Truth serums perform better in predicting own choices than introspection. On the other hand, they can lead to even larger deviations from additivity than introspection does. Directly providing a probability estimate may help participants to adhere to the basic additivity principles, compared to the revealed preference measures in which probabilities are not explicitly specified. We find evidence that correcting for nonlinear risk attitudes improves the additivity of beliefs elicited by the QSR.

External Validity of the Elicited Beliefs. The third quality criterion that we employ concerns the external validity of the elicited beliefs. We assess the consistency of the beliefs with the actual observed choice/acceptance probabilities. Table 5 shows the observed acceptance frequencies of responders for each proposal, as well as the reported beliefs of the proposers in the different treatments. The subjective probabilities of the proposers appear to be systematically distorted towards uniform beliefs in almost all treatments. That is, for proposals that are rarely accepted, proposers are too optimistic; for proposals that are likely to be accepted, proposers are too pessimistic. A similar pattern in reported beliefs was found by Costa-Gomez and Weizsäcker (2008) and Bellemare et al. (2009), and is consistent with *conservatism* in reported beliefs (Edwards 1954). It is also consistent with the tendency for agents to overweight low probabilities and underweight large probabilities, i.e., with an

inverse-S shaped probability weighting function (Tversky and Kahneman 1992; Carman and Kooreman 2010).

Table 5 also shows the effect of the correction methods for the QSR and outcome matching. For the QSR, risk aversion biases reported beliefs toward uniformity as discussed in section 3. The correction leads to significantly less uniformity (comparison column 6 and 8, $p<.05$, Wilcoxon test). For outcome matching, concave utility leads to a downward bias in reported beliefs.[14] Comparison of columns 4 and 7 of Table 5 shows that correction increases the beliefs ($p<.05$, Wilcoxon test). However, the uniformity bias remains strong for corrected outcome matching.

TABLE 5: Observed Acceptance Frequencies versus Proposer's Beliefs

Allo-cation	Observed Frequency	Intro-spection	Outcome Matching	Probability Matching	QSR	Outcome Matching (corrected)	QSR (corrected)
(€20, €0)	14	42***	50***	44***	41***	52***	39***
		(36)	(22)	(22)	(26)	(18)	(30)
(€16, €4)	43	58**	58**	45	44	60***	40
		(28)	(27)	(22)	(18)	(22)	(24)
(€12, €8)	90	71***	65***	62***	67***	66***	70***
		(25)	(27)	(20)	(16)	(18)	(21)
(€8, €12)	95	82**	73***	66***	72***	74***	76***
		(19)	(23)	(19)	(21)	(20)	(23)
(€4, €16)	92	87	75***	71***	73***	75***	80**
		(19)	(26)	(21)	(19)	(23)	(20)
(€0, €20)	88	94***	73	71**	78	74	82
		(17)	(33)	(30)	(24)	(30)	(27)

Notes: Numbers are percentages; */**/*** denotes significant different from observed frequencies based on a two-sided Wilcoxon signed-rank test at the 10%/5%/1% Level. Standard deviations in parenthesis.

As a measure of accuracy of reported beliefs, we calculate mean *Brier scores*, the average squared deviation between the actual choice frequencies and the reported beliefs, for each of the six allocations over all proposers (Costa-Gomez and Weizsaecker 2008). Lower scores indicate higher levels of accuracy. Results are in Table 6. All methods perform better than a random prediction derived from uniform distribution between 0 and 100.[15] Among the

[14] In B(E)=(CE-y)/(x-y), for concave utility, the difference x-y is overestimated compared to U(x) – U(y) more strongly than the difference CE-y is overestimated compared to U(CE)-U(y).

[15] Scores for random prediction calculated as $\frac{1}{101}\sum_{n=0}^{100}(true\ prob - n)^2$.

different elicitation methods, however, there are no systematic differences. In particular, introspective beliefs are as accurate as incentivized beliefs.

TABLE 6: Mean Brier Score of Reported Beliefs by Proposers

Allo-cation	Random prediction[a]	Intro-spection	Outcome Matching	Probability Matching	QSR	Outcome Matching (corrected)	QSR (corrected)
(€20, €0)	.219	0.205	0.177	0.141	0.138	0.176	0.151
(€16, €4)	.089	0.098	0.089	0.046	0.033	0.076	0.057
(€12, €8)	.241	0.096	0.115	0.121	0.080	0.091	0.084
(€8, €12)	.282	0.052	0.100	0.116	0.094	0.081	0.089
(€4, €16)	.257	0.035	0.098	0.086	0.070	0.079	0.052
(€0, €20)	.226	0.033	0.128	0.115	0.066	0.106	0.073
Mean	0.219	0.087*	0.118*	0.104*	0.080*	0.102*	0.084*

Notes: Lower scores represent higher levels of accuracy. a: for each event the belief is calculated as the random prediction from a uniform distribution *: significantly different form random prediction at the 5% significance level, Mann-Whitney U test

For responders, Table 7 and 8 show the results of the analogous analyses. There is conservatism in the beliefs with a tendency toward uniform beliefs. The correction methods affect beliefs in the same direction as discussed above for proposers. Regarding the accuracy, there are no significant differences in the mean Brier scores between any of the methods, with all mean scores very similar. All methods improve accuracy compared to the benchmark of random predictions.

Interestingly, although the correction for the QSR reduces the risk aversion bias and increases average accuracy for all events (Table 5 and 7), variances of the corrected beliefs, and also brier scores, go up. In Appendix C we have shown that for known probabilities, subjects are biased toward probability .5, but not beyond the middle of the table. For unknown probabilities, however, subjects may be less well calibrated, holding beliefs lower than .5 for an event that occurs with a probability above .5 for instance. In this situation, correction for risk aversion will further move the corrected beliefs away from the objective probabilities, reducing accuracy. That is, eliminating the risk aversion bias in QSR beliefs does not necessarily increase accuracy.

TABLE 7: Observed Choice Frequencies versus Responder's Beliefs

Allo-cation	Observed Frequency	Intro-spection	Outcome Matching	Probability Matching	QSR	Outcome Matching (corrected)	QSR (corrected)
(€20, €0)	6	27 (14)	33*** (14)	39*** (25)	39*** (24)	36*** (18)	34*** (27)
(€16, €4)	20	48*** (29)	41*** (14)	46*** (21)	54*** (26)	44*** (16)	47*** (28)
(€12, €8)	66	54** (29)	45*** (13)	45*** (22)	62 (18)	47*** (16)	63 (20)
(€8, €12)	7	35*** (30)	38*** (14)	30*** (23)	43*** (13)	40*** (21)	37*** (16)
(€4, €16)	2	16** (21)	28*** (17)	16*** (16)	28*** (22)	30*** (23)	20*** (21)
(€0, €20)	0	2** (4)	25*** (21)	25*** (19)	11*** (21)	28*** (23)	8 (21)

Notes: Numbers are percentages; ***/**/* denotes significant different from observed frequencies based on a two-sided Wilcoxon signed-rank test, at the 1%/5%/10% level. Standard deviations in parenthesis.

TABLE 8: Mean Brier Score of Reported Beliefs by Responders

Allo-cation	Random prediction[a]	Intro-spection	Outcome Matching	Probability Matching	QSR	Outcome Matching (corrected)	QSR (corrected)
(€20, €0)	0.282	0.196	0.094	0.167	0.163	0.122	0.144
(€16, €4)	0.179	0.160	0.066	0.113	0.181	0.085	0.152
(€12, €8)	0.106	0.095	0.061	0.091	0.033	0.059	0.040
(€8, €12)	0.274	0.168	0.123	0.102	0.146	0.150	0.113
(€4, €16)	0.319	0.063	0.111	0.042	0.113	0.132	0.074
(€0, €20)	0.338	0.002	0.106	0.096	0.054	0.131	0.050
Mean	0.250	0.114*	0.093*	0.102*	0.115*	0.113*	0.096*

Notes: Lower scores represent higher levels of accuracy. a: for each event the belief is calculated as the random prediction from a uniform distribution *: significantly different form random prediction at the 5% significance level, Mann-Whitney U test.

6. Discussion

This paper compares the performance of introspective belief measurement to five incentive compatible elicitation methods, or truth serums, of different degree of complexity. More complex methods are more difficult to implement in survey studies, economic experiments,

or decision analyses. A researcher who is constrained in terms of time, number of questions asked, or funds for monetary incentives therefore needs to make an informed trade-off between the cost and the benefits of complexity. The current study includes three widely used truth serums, and two corrected methods that account for deviations from risk neutrality. Importantly, we consider a larger range of quality criteria to rank methods than previous studies did. We test both the internal validity (additivity, prediction of own behavior), and the external validity (accuracy of other players' behavior) of the elicited beliefs. Our findings suggest that there is no clear benefit from using more complex methods. First, there are virtually no differences in the performance of the truth serums for any of the quality criteria. Second, comparison of the incentivized methods with simple introspection does not reveal a clear advantage for the truth serums.

Accuracy. In terms of accuracy of the beliefs, the introspective method provides equally good predictions as the incentive compatible methods. All methods generally perform better than random prediction. The evidence regarding the accuracy of introspective versus incentivized beliefs is mixed in the literature (Friedman and Massaro 1998; Hollard et al. 2010; Sonnemans and Offerman 2001; Rutstroem and Wilcox 2009). The current data support the view that there are no clear differences in accuracy between introspection and incentivized methods.

A potential explanation for the observed differences in the literature lies in the presence or absence of an underlying social decision task. Non-incentivized methods may be affected by misrepresentation of beliefs, in contrast to simple lack of effort and increased noise due to the lack of incentives. Beliefs in social settings may be more prone to, possibly unconscious, justification and accountability pressures (Vieider 2011; Vieider and Tetlock 2010; Zizzo 2010). Indeed, studies that elicited beliefs about events in individual decision tasks found no evidence of a bias in introspective beliefs (Friedman and Massaro 1998; Hollard et al. 2010; Sonnemans and Offerman 2001), while Rutstroem and Wilcox (2009) report worse performance of introspection than the QSR in a task based on the prediction of another player's behavior. The current study involved beliefs in an ultimatum game—a strategic setting that has often been associated with justification and strategic misrepresentation of attitudes and beliefs—and finds similar accuracy among incentivized and non-incentivized methods. Thus, intentional misrepresentation does not seem to be a omnipresent problem in belief elicitation in social settings.

A few studies have pointed out that uncorrected outcome matching and scoring rule methods are not incentive compatible in the presence of deviations from risk neutrality, and may thus fail to elicit true beliefs (Andersen et al. 2010; Heineman et al. 2009; Hollard et al. 2010; Hao and Houser 2010, Offerman et al 2009; Sonnemans and Offerman 2001). We find evidence for deviations caused by deviations from risk neutrality, and find that the methods proposed to correct these methods work in the intended direction: for outcome matching, the downward bias is reduced, and for the QSR the bias toward uniform beliefs is reduced by using corrected measures. However, these corrections did not lead to an overall significantly increased accuracy of beliefs.

Predicting behavior. In predicting proposers' own choices from their beliefs, truth serums improve on random prediction while introspection does not. This holds true for different specifications of the (social) utility function. The predictive power of all methods is at best modest, however, and objective probabilities clearly outperform subjective beliefs. Our results are consistent with the previous findings of modest internal validity of subjective beliefs (Costa-Gomes and Weizsaecker 2008), although the poor performance of introspection in predicting choice behavior is somewhat surprising. There is evidence showing that non-incentivized introspective beliefs predict behavior in many settings (Bosman and van Winden 2002; Carman and Kooreman 2010; Guiso et al. 1992; Guiso and Parigi 1999). However, the link between subjective probabilities and behavior might be weaker in strategic interaction task compared to individual decisions. In the current setting, unobserved social preferences that are not properly modeled by the expected (social) utility function may play an important role in the decision process. For instance, the favorable allocations (€4, €16) and especially (€0, €20) are rejected by roughly 10% of the responders. Such behavior cannot be explained by the models that we considered. It is therefore conceivable that in situations that do not involve strong social aspects, e.g. the medical decisions studied in Carman and Kooreman (2010) or the financial decisions studied by Guiso et al (1992) and Guiso and Parigi (1999), the predictive power of subjective beliefs is higher than in the current task.

Additivity. Violations of additivity have often been observed for subjective beliefs, and Tversky and Koehler (1994) offer a psychological account of such violations. Consistent with their theory, we find that violations of additivity are more severe for 6 events compared to 2 events. Tversky and Koehler (1994) conjecture that non-additivity may be a problem

particularly related to introspective beliefs.[16] However, Offermann et al. (2009) find non-additivity for the QSR and the corrected QSR. The current paper shows that non-additivity is strong for all five truth serums, and that introspection sometimes outperforms incentive compatible methods. Correcting the scoring rule beliefs for risk attitudes reduces deviations from additivity, but the absolute level of deviation from additivity remains large.

7. Conclusion

Subjective beliefs are an important input in many economic policy and management decisions, ranging from expectations about macroeconomic variables to individual epidemiological risk factors in medical decisions. The most commonly used method to elicit subjective beliefs outside experimental settings and decision analyses involves the simple reporting of introspective beliefs (Delavande 2011, Hurd 2009). The reason for the popularity of the non-incentivized introspective method lies in its simplicity and the potential difficulty of implementing incentive compatible payments on survey panels and for events that are unobservable during the time frame of the study. In experiments, simple methods are often preferred because they potentially distract subjects from the main task of interest (Cabrales et al 2010; Haruvy et al. 2007). Our study compared simple introspection to five incentive compatible belief elicitation methods of different complexity, where more complexity implies theoretically more robust methods. We compare the different methods in a controlled setting with uncertain events with probabilities unknown to the subjects, but observable by the researcher. That allows us to study both internal and external validity in an environment of natural uncertainty. We found little evidence for improved performance of more complex truth serums. Moreover, the results show that introspective beliefs measurement performs similar in terms of accuracy and additivity as the incentivized methods. This finding supports the view that introspection is a valid method to measure subjective beliefs in many applications. However, incentivized beliefs are a better predictor of players' own behavior in the game. Incentivized methods may also reduce the risk of participants knowingly or unknowingly misrepresenting their beliefs. On balance, thus, our findings imply that if complexity is feasible, like in laboratory experiments or expert decision analysis, researchers can benefit from using more complex and incentivized methods. For many purposes, however, introspection will serve the researcher just as well.

[16] See also Attanasio (2009).

References

Abdellaoui, Mohammed, Frank Vossmann, and Martin Weber (2005). Choice-Based Elicitation and Decomposition of Decision Weights for Gains and Losses under Uncertainty. *Management Science* 51, 1384–1399.

Andersen, Steffen, John Fountain, Glenn W. Harrison, and Elisabet Rutstroem (2010). Estimating Subjective Probabilities. Working paper, CEAR.

Armantier, Olivier and Nicolas Treich (2009). Subjective Probabilities in Games: An Application to the Overbidding Puzzle. *International Economic Review* 50, 1079–1102.

Arrow, Kenneth J. (1951). Alternative Approaches to the Theory of Choice in Risk-Taking Situations. *Econometrica* 19, 404–437.

Attanasio, Orazio (2009). Expectations and Perceptions in Developing Countries: Their Measurement and Their Use. *American Economic Review* 99, 87–92.

Bellemare, Charles, Sabine Kroger, and Arthur van Soest (2008). Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities. *Econometrica* 76, 815–839.

Blanco, Mariana, Dirk Engelmann, Alexander K. Koch, and Hans-Theo Norman (2010). Belief Elicitation in Experiments: Is There a Hedging Problem? *Experimental Economics*, 13, 412–438.

Blanco, Mariana, Dirk Engelmann, Alexander K. Koch, and Hans-Theo Norman (2011). Preferences and Beliefs in a Sequential Social Dilemma: A Within-Subject Analysis. Working paper, Mannheim University.

Bosman, Ronald, and Frans van Winden (2002). Emotional Hazard in a Power-to-Take Experiment. *Economic Journal* 112, 147–169.

Branch, William A. (2004). The Theory of Rationally Heterogeneous Expectations: Evidence from Survey Data on Inflation Expectations. *Economic Journal* 114, 592–621.

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

Cabrales, Antonio, Raffaele Miniaci, Marco Piovesan, and Giovanni Ponti (2010). Social Preferences and Strategic Uncertainty: An Experiment on Markets and Contracts. *American Economic Review* 100, 2261–2278.

Cameron, Lisa , and Manisha Shah (2011). Risk-Taking in the Wake of Natural Disasters. Working paper, Monash.

Carman, Katie G., and Peter Kooreman (2010). Flu Shots, Mammogram, and the Perception of Probabilities. Netspar working paper 2010-14.

Costa-Gomes, Miguel, and Georg Weizsäcker (2008). Stated Beliefs and Play in Normal-Form Games. *Review of Economic Studies* 75, 729–762.

Costa-Gomes, Miguel, Steffen Huck, and Georg Weizsäcker (2008). Beliefs and Actions in the Trust Game: Creating Instrumental Variables to Estimate the Causal Effect. Working paper, DIW Berlin.

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

Delavande Adeline, Xavier Gine, and David McKenzie (2011). Measuring subjective expectations in developing countries: A critical review and new evidence. *Journal of Development Economics*, 151–163.

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.

Fehr, Ernst and Klaus Schmidt (1999). A theory of fairness, competition and cooperation. *Quarterly Journal of Economics*, 114, 817–868.

Fehr, Ernst, and Klaus Schmidt (2003). Theories of Fairness and Reciprocity - Evidence and Economic Applications. In: M. Dewatripont, L. Hansen and St. Turnovsky (Eds.), *Advances in Economics and Econometrics - 8th World Congress, Econometric Society Monographs,* Cambridge, Cambridge University Press, 208–257.

Fischbacher, Urs (2007). Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics* 10, 171–178.

Friedman, Daniel, and Dominic W. Massaro (1998). Understanding Variability in Binary and Continuous Choice. *Psychonomic Bulletin & Review* 5, 370–389.

Guiso, Luigi, Tullio Jappelli, and Daniele Terlizzese (1992). Earnings Uncertainty and Precautionary Saving. *Journal of Monetary Economics* 30, 307–337.

Guiso, Luigi, and Giuseppe Parigi (1999). Investment and Demand Uncertainty. *Quarterly Journal of Economics* 114, 185–227.

Hao, Li, and Daniel Houser (2010). Getting it right the first time: Belief elicitation with novice participants. Working paper, George Mason University.

Haruvy, Ernan, Yaron Lahav, and Charles N. Noussair, C. N. (2007). Traders' Expectations in Asset Markets: Experimental Evidence. *American Economic Review* 97, 1901–1920.

Heinemann, Frank, Rosemarie Nagel, and Peter Ockenfels (2009). Measuring Strategic Uncertainty in Coordination Games. *Review of Economic Studies* 76, 181–221.

Hennig-Schmidt, Heike, Zhu-Yu Li, and Chaoliang Yang (2008). Why People Reject Advantageous Offers: Non-Monotonic Strategies in Ultimatum Bargaining. *Journal of Economic Behavior and Organization* 65, 373–384.

Hollard, Guillaume, Sebastien Massoni, and Jean-Christophe Vergnaud (2010). Subjective belief formation and elicitation rules: Experimental evidence. Working paper, CES Paris.

Holt, Charles A. (2006). *Webgames and Strategy: Recipes for Interactive Learning*. Addison-Wesley, forthcoming.

Hossain, Tanjim and Ryo Okui (2011). The Binarized Scoring Rule. Working paper, Toronto.

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

Hurd, Michael D. (2009). Subjective Probabilities in Household Surveys. Annual Reviews of Economics 1, 543–562.

Hurwicz, Leonid (1960). Optimality and Informational Efficiency in Resource Allocation. In: Kenneth J. Arrow, Samuel Karlin, and Patrick Suppes (1960, Eds), *Mathematical Methods in the Social Sciences*, 17–46, Stanford University Press, Stanford, CA.

Kadane, Joseph B., and Robert L. Winkler (1988). Separating Probability Elicitation from Utilities. *Journal of the American Statistical Association* 83, 357–363.

Karni, Edi (2009). A Mechanism Design for Probability Elicitation. *Econometrica* 77, 603 – 606.

Khwaja, Ahmed, Dan Silverman, Frank.A. Sloan, and Yang Wang (2007). Are Mature Smokers Misinformed? *Journal of Health Economics* 28, 385–397.

Khwaja, Ahmed, Frank Sloan, and Martin Salm (2006) Evidence on Preferences and Subjective Beliefs of Risk Takers: The Case of Smokers. *International Journal of Industrial Organization* 24, 667–682.

Kothiyal, Amit, Vitali Spinu, and Peter P. Wakker (2011). Comonotonic Proper Scoring Rules to Measure Ambiguity and Subjective Beliefs. *Journal of Multi-Criteria Decision Analysis*, forthcoming.

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, 1175–1194.

Machina, Mark J. (1987). Decision-Making in the Presence of Risk. *Science* 236, 537–543.

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

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

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

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

Offerman, Theo, Joep Sonnemans, Gijs van de Kuilen, and Peter P. Wakker (2009). A Truth-Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes. *Review of Economic Studies* 76, 1461–1489.

Oosterbeek, Hessel, Randolph Sloof, and Gijs van de Kuilen (2004). Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis. *Experimental Economics* 7, 171–188.

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

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

Rey-Biel, Pedro (2009). Equilibrium play and best response to (stated) beliefsin normal form games. Games and Economic Behavior 65, 572–585.

Roth, Al (1995). Bargaining Experiments. In: J. Kagel and A. Roth (eds.), *The Handbook of Experimental Economics*. Princeton: Princeton University Press, 3–109.

Ruthstrom, E. Elisabeth and Nathaniel T. Wilcox (2009). Stated Beliefs versus Inferred Beliefs: A Methodological Inquiry and Experimental Test. *Games and Economic Behavior* 67, 616–632.

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

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

Sonnemans, Joep and Theo Offerman (2001). Is the quadratic scoring rule really incentive compatible? Working paper, University of Amsterdam.

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

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

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

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

Vieider, Ferdinand M. (2011). Separating Real Incentives and Accountability. *Experimental Economics*, forthcoming.

Vieider, Ferdinand M. and Philip E. Tetlock (2010). Accountability: A Meta-Analysis of Effect Sizes and Situated Identity Analysis of Research Settings. Working Paper, UPenn.

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

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

Zizzo, Daniel J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics* 13, 75–98.

Appendix (not for publication; will be made available online)

A Example Instructions

This appendix provides the instructions for the ultimatum game task and for the different belief elicitation methods (between-subject). We give for each method instructions for either the responder or the proposer. Screenshots of the choice lists are given in appendix B. In the experiment, instructions were given on the left hand side of the screen, while decision forms and choice lists were given on the right hand side of the screen.

Ultimatum game task, general

In this experiment, participants are randomly assigned the role of proposer or responder. In the first period,[17] proposers are asked to divide €20 between themselves and a responder, by choosing 1 of the 6 possible allocations depicted in the table below:

	Allocation 1	Allocation 2	Allocation 3	Allocation 4	Allocation 5	Allocation 6
Proposer	€20	€16	€12	€8	€6	€4
Responder	€0	€4	€8	€12	€16	€20

At the same time, responders are asked whether they would accept or reject each of the 6 possible allocations. At the end of the experiment, each proposer is randomly matched with a responder. The earnings of proposers and responders in the first period are determined by whether or not the responder accepted the allocation chosen by the proposer.

If the responder accepted the allocation chosen by the proposer, the €20 is divided between the proposer and the responder in accordance with the chosen allocation. If the responder rejected the allocation chosen by the proposer, the earnings of the responder and proposer in the first period are equal to €0.

Ultimatum game task, proposer

You are a proposer. On the right, you are asked to choose an allocation by typing in a number from 1 till 6. At the end of the experiment, the computer will select one period at random to be paid for real. If the computer then selects period 1, your earnings are determined as follows. First, you will be randomly matched with a responder. This responder was asked to accept or reject each of the 6 possible allocations in period 1. If the responder that is matched to you rejected the allocation you have chosen in period 1, you and the responder will receive €0. If

[17] In the experiment we used the term period instead of stage. Each decision screen was a period, that is, there were more periods than stages as defined in the design section because stage 2 contained 7 different belief questions.

the responder that is matched to you accepted the allocation you have chosen in period 1, the €20 will be divided between you and the responder in accordance with the allocation chosen by you.
For instance, if the computer selected period 1 to be paid for real, you chose Allocation 4 in period 1, and the responder matched to you accepted Allocation 4 in period 1, you will get €8. In this case, the responder matched to you will get €12.
Please choose an allocation by typing in a number from 1 till 6 on the right side of the screen.

Ultimatum game task, responder
You are a responder. On the right, you are asked to reject or accept each of the 6 possible allocations. At the end of the experiment, the computer will one period at random to be paid for real. If the computer then selects period 1, your earnings are determined as follows. First, you will be randomly matched with a proposer. This proposer was asked to choose 1 of the 6 possible allocations in Period 1. If you have rejected the allocation chosen by the proposer in period 1, you and the proposer matched to you will receive nothing. If you have accepted the allocation chosen by the proposer in Period 1, the € 20 will be divided between you and the proposer in accordance with the allocation chosen by the proposer.
For instance, if the computer selected period 1 to be paid for real, the proposer matched to you chose Allocation 4 in Period 1, and you accepted Allocation 4 in period 1, you will get €12. In this case, the responder matched to you will get €8.
Please reject or accept each of the 6 possible allocations, by ticking the "reject" or "accept" button for each allocation on the right side of the screen.

Belief task, proposer, introspection
You are a proposer. At the end of the experiment, you will be randomly matched with a responder. This responder was asked to accept or reject each of the 6 possible allocations in period 1.
On the right, we ask you to report the probability that you think that the responder matched to you accepted Allocation 1 in period 1. We ask you to report this probability in percentages, ranging from 0% to 100%. For example, if you are completely sure the responder matched to you at the end of the experiment accepted Allocation 1 in period 1, you should report a probability of 100%. If you are not sure whether the responder matched to you accepted Allocation 1, you should report a probability between 0% and 100%.
At the end of the experiment, the computer will select one period at random to be paid for real. If the computer then selects period 2, your earnings are €5.

Belief task, proposer, probability matching
You are a proposer. At the end of the experiment, you will be randomly matched with a responder. This responder was asked to accept or reject each of the 6 possible allocations in period 1.
On the right, you see a list of choices between options labeled Option A, and Option B. In each choice, Option A yields Asset [asset number], depicted below:

Asset [asset number]

If the responder matched to you accepted allocation [alloc. number]:	€15
If the responder matched to you rejected allocation [alloc. number]:	€0

Thus, Asset [asset number] yields €15 if the responder matched to you at the end of the experiment accepted Allocation [alloc. number] in the first period. If the responder matched to you at the end of the experiment rejected Allocation [alloc. number] in the first period, Asset [asset number] yields €0.

In each choice, Option B yields an amount of money depending on the roll of a 20-sided die. To determine the amount of money that Option B yields, a 20-sided die will be rolled at the end of the experiment. For instance, Option B in Choice 4 yields €15 if the roll with the 20-sided die is 1, 2, or 3. Otherwise, Option B in Choice 4 yields €0.

Now please take a look at Choice 1 in the list of choices on the right. We imagine that most people would choose Option A in Choice 1, since Option A then gives a chance of an amount higher than zero, whereas Option B gives €0 for sure. Similarly, we imagine that most people would choose Option B in Choice 21, since Option B then gives €15 for sure, whereas Option A only gives a chance of €15. Hence, we imagine that most people would switch from choosing Option A to Option B at some point in the list.

You are asked to make 21 choices between Option A and Option B by ticking the box corresponding with the option you prefer. Although we imagine that most people would switch from Option A to Option B at some point in the list, it is entirely up to you what to do in each of the choices. At the end of the experiment, the computer will select one period at random to be paid for real. If the computer then selects period [current period], your earnings are determined as follows.

First, you will be randomly matched with a responder, and the computer will select 1 of the 21 choices at random. The option you have chosen in that choice will then be paid out for real, depending on the roll of the 20-sided die (if you have chosen Option B) or the decision made by the responder matched to you (if you have chosen Option A). Thus, each of your choices could eventually be the one that determines the payment you receive.

Belief task, responder, outcome matching

You are a responder. At the end of the experiment, you will be randomly matched with a proposer. This proposer was asked to choose 1 of the 6 possible allocations in period 1. On the right, you see a list of choices between options labeled Option A, and Option B. In each choice, Option A yields Asset [asset number], depicted below:

Asset [asset number]

If the responder matched to you did choose allocation [alloc. number]:	€15
If the responder matched to you did not choose allocation [alloc. number]:	€0

Thus, Asset [asset number] yields €15 if the proposer matched to you did choose Allocation [alloc. number] in the first period. If the proposer matched to you did not choose Allocation [alloc. number] in the first period, Asset [asset number] yields €0. In each choice, Option B yields a certain amount of money.

Now please take a look at Choice 1 in the list of choices on the right. We imagine that most people would choose Option A in Choice 1, since Option A then gives a chance of an amount higher than €0, whereas Option B gives €0 for sure. Similarly, we imagine that most people would choose Option B in Choice 21, since Option B then gives €15 for sure, whereas Option A only gives a chance of €15. Hence, we imagine that most people would switch from choosing Option A to Option B at some point in the list. You are asked to make 21 choices between Option A and Option B by ticking the box corresponding with the option you prefer. Although we imagine that most people would switch from Option A to Option B at some point in the list, it is entirely up to you what to do in each of the choices. At the end of the experiment, the computer will select one period at random to be paid for real. If the computer then selects period [current period], your earnings are determined as follows. First, you will be randomly matched with a proposer, and the computer will select 1 of the 21 choices at random. The option you have chosen in that choice will then be paid out for real, depending on the decision made by the proposer matched to you in case you have chosen Option A in that choice. Thus, each of your choices could prove to be the one that determines the payment you receive.

Belief task, responder, QSR

You are a responder. At the end of the experiment, you will be randomly matched with a proposer. This proposer was asked to choose one of the six possible allocations in period 1.
On the right, you see a Decision Table with 21 numbered rows.[18] Each row yields an amount of euros, depending on whether the proposer matched to you did or did not chose Allocation 1 in the first period. The amount of euros that each row yields if the proposer matched to you did choose Allocation 1 is shown in the second column of the Decision Table. The amount of euro that each row yields if the proposer matched to you did not choose Allocation 1 is shown in the third column of the Decision Table.
For instance, if the responder matched to you at the end of the experiment did choose Allocation 1 in the first period, Row 5 yields €7.20. If the responder matched to you at the end of the experiment did not choose Allocation 1 in the first period, Row 5 yields €19.20.
You are asked to choose a row by typing in the number of the row of your choice in the box below the Decision Table. At the end of the experiment, the computer will select one of the periods at random to be paid for real. If Period [current period] then is selected, you will receive the amount of euro corresponding to the row that you have chosen, depending on whether the proposer matched to you did or did not choose Allocation 1. Thus, each of your choices could prove to be the one that determines the payment you receive.

Risk aversion task
On the right, you see a list of choices between options labeled Option A and Option B. In each choice, Option A yields Asset [asset number], depicted below:

Asset [asset number]

if roll of 20-sided die is 1-10:	€10
if roll of 20-sided die is 11-20:	€0

[18] See screen shot example in Appendix B.

Thus, Asset [asset number] yields €10 if the roll of a 20-sided die is 1 till 10 (1, 2, 3, 4, 5, 6, 7, 8, 9, or 10). If the roll of a 20-sided die is 11 till 20 (11, 12, 13, 14, 15, 16, 17, 18, 19, or 20), Asset [asset number] yields €0. To determine the amount of money that Asset [asset number] yields, a 20-sided die will be rolled at the end of the experiment. In each choice, Option B yields a certain amount of money.

Now please take a look at Choice 1 in the list of choices on the right. We imagine that most people would choose Option A in Choice 1, since Option A then gives a chance of an amount higher than €0, whereas Option B gives €0 for sure. Similarly, we imagine that most people would choose Option B in Choice 21, since Option B then gives €10 for sure, whereas Option A only gives a chance of €10. Hence, we imagine that most people would switch from choosing Option A to Option B at some point in the list.

You are asked to make 21 choices between Option A and Option B by ticking the box corresponding with the option you prefer. Although we imagine that most people would switch from Option A to Option B at some point in the list, it is entirely up to you what to do in each of the choices.

At the end of the experiment, the computer will select one period at random to be paid for real. If the computer then selects period [current period], your earnings are determined as follows. First, the computer will select 1 of the 21 choices at random. The option you have chosen in that choice will then be paid out for real, depending on the roll of the 20-sided die in case you have chosen Option A in that choice. Thus, each of your choices could prove to be the one that determines the payment you receive.

B Example Screen Shots

Screenshot, table outcome matching

Choice	Option A	Option B	Your Choice:
1	ASSET 1	€ 0.00	Option A ○ ○ Option B
2	ASSET 1	€ 0.75	Option A ○ ○ Option B
3	ASSET 1	€ 1.50	Option A ○ ○ Option B
4	ASSET 1	€ 2.25	Option A ○ ○ Option B
5	ASSET 1	€ 3.00	Option A ○ ○ Option B
6	ASSET 1	€ 3.75	Option A ○ ○ Option B
7	ASSET 1	€ 4.50	Option A ○ ○ Option B
8	ASSET 1	€ 5.25	Option A ○ ○ Option B
9	ASSET 1	€ 6.00	Option A ○ ○ Option B
10	ASSET 1	€ 6.75	Option A ○ ○ Option B
11	ASSET 1	€ 7.50	Option A ○ ○ Option B
12	ASSET 1	€ 8.25	Option A ○ ○ Option B
13	ASSET 1	€ 9.00	Option A ○ ○ Option B
14	ASSET 1	€ 9.75	Option A ○ ○ Option B
15	ASSET 1	€ 10.50	Option A ○ ○ Option B
16	ASSET 1	€ 11.25	Option A ○ ○ Option B
17	ASSET 1	€ 12.00	Option A ○ ○ Option B
18	ASSET 1	€ 12.75	Option A ○ ○ Option B
19	ASSET 1	€ 13.50	Option A ○ ○ Option B
20	ASSET 1	€ 14.25	Option A ○ ○ Option B
21	ASSET 1	€ 15.00	Option A ○ ○ Option B

Screen shot, table probability matching

Choice	Option A	Option B a 20-sided die is rolled at the end of the experiment:	Your Choice:
1	ASSET 1	roll is 1-20: €0.00	Option A ○ ○ Option B
2	ASSET 1	roll is 1: €15.00 roll is 2-20: €0.00	Option A ○ ○ Option B
3	ASSET 1	roll is 1-2 : €15.00 roll is 3-20: €0.00	Option A ○ ○ Option B
4	ASSET 1	roll is 1-3 : €15.00 roll is 4-20: €0.00	Option A ○ ○ Option B
5	ASSET 1	roll is 1-4 : €15.00 roll is 5-20: €0.00	Option A ○ ○ Option B
6	ASSET 1	roll is 1-5 : €15.00 roll is 6-20: €0.00	Option A ○ ○ Option B
7	ASSET 1	roll is 1-6 : €15.00 roll is 7-20: €0.00	Option A ○ ○ Option B
8	ASSET 1	roll is 1-7 : €15.00 roll is 8-20: €0.00	Option A ○ ○ Option B
9	ASSET 1	roll is 1-8 : €15.00 roll is 9-20: €0.00	Option A ○ ○ Option B
10	ASSET 1	roll is 1-9 : €15.00 roll is 10-20: €0.00	Option A ○ ○ Option B
11	ASSET 1	roll is 1-10 : €15.00 roll is 11-20: €0.00	Option A ○ ○ Option B
12	ASSET 1	roll is 1-11 : €15.00 roll is 12-20: €0.00	Option A ○ ○ Option B
13	ASSET 1	roll is 1-12 : €15.00 roll is 13-20: €0.00	Option A ○ ○ Option B
14	ASSET 1	roll is 1-13 : €15.00 roll is 14-20: €0.00	Option A ○ ○ Option B
15	ASSET 1	roll is 1-14 : €15.00 roll is 15-20: €0.00	Option A ○ ○ Option B
16	ASSET 1	roll is 1-15 : €15.00 roll is 16-20: €0.00	Option A ○ ○ Option B
17	ASSET 1	roll is 1-16 : €15.00 roll is 17-20: €0.00	Option A ○ ○ Option B
18	ASSET 1	roll is 1-17 : €15.00 roll is 18-20: €0.00	Option A ○ ○ Option B
19	ASSET 1	roll is 1-18 : €15.00 roll is 19-20: €0.00	Option A ○ ○ Option B
20	ASSET 1	roll is 1-19 : €15.00 roll is 20: €0.00	Option A ○ ○ Option B
21	ASSET 1	roll is 1-20: €15.00	Option A ○ ○ Option B

Screen shot, table QSR

Decision Table

Row #	Payoff if responder ACCEPTED Allocation 1	Payoff if responder REJECTED Allocation 1
1	€ 0.00	€ 20.00
2	€ 1.95	€ 19.95
3	€ 3.80	€ 19.80
4	€ 5.55	€ 19.55
5	€ 7.20	€ 19.20
6	€ 8.75	€ 18.75
7	€ 10.20	€ 18.20
8	€ 11.55	€ 17.55
9	€ 12.80	€ 16.80
10	€ 13.95	€ 15.95
11	€ 15.00	€ 15.00
12	€ 15.95	€ 13.95
13	€ 16.80	€ 12.80
14	€ 17.55	€ 11.55
15	€ 18.20	€ 10.20
16	€ 18.75	€ 8.75
17	€ 19.20	€ 7.20
18	€ 19.55	€ 5.55
19	€ 19.80	€ 3.80
20	€ 19.95	€ 1.95
21	€ 20.00	€ 0.00

Choose a row:

(Type in a number from 1 till 21)

C Risk Correction of Scoring Rule Beliefs

Figure C1 illustrates the effect of risk attitude on beliefs elicited by the uncorrected QSR. The figure plots the median reported probabilities, implied by the scoring rule choices, against the true probabilities of the known risks. All deviations bias the reported beliefs toward .5. All deviations are significantly different than the true probability except for p=.5 (two-sided Wilcoxon tests).

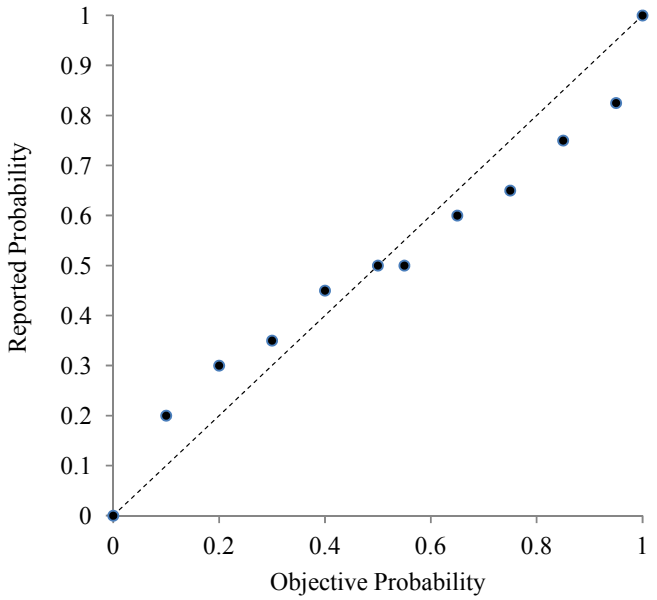

Figure C1: Objective Probabilities versus QSR-implied Probabilities for Known Risks