J Risk Uncertain (2014) 48:207–229
DOI 10.1007/s11166-014-9194-z

Estimating subjective probabilities

Steffen Andersen · John Fountain · Glenn W. Harrison · E. Elisabet Rutström

Published online: 12 July 2014

Abstract Subjective probabilities play a central role in many economic decisions and act as an immediate confound of inferences about behavior, unless controlled for. Several procedures to recover subjective probabilities have been proposed, but in order to recover the correct latent probability one must either construct elicitation mechanisms that control for risk aversion, or construct elicitation mechanisms which undertake "calibrating adjustments" to elicited reports. We illustrate how the *joint estimation of risk attitudes and subjective probabilities* can provide the calibration adjustments that theory calls for. We illustrate this approach using data from a controlled experiment with real monetary consequences to the subjects. This allows the observer to make inferences about the latent subjective probability, under virtually any well-specified model of choice under subjective risk, while still employing relatively simple elicitation mechanisms.

JEL Classifications C9 · C8 · D84

Electronic supplementary material The online version of this article (doi:10.1007/s11166-014-9194-z) contains supplementary material, which is available to authorized users.

S. Andersen
Department of Economics, Copenhagen Business School, Copenhagen, Denmark

J. Fountain
CEAR, Robinson College of Business, Georgia State University, Atlanta, GA 30303, USA

G. W. Harrison
Department of Risk Management & Insurance and CEAR, Robinson College of Business, Georgia State University, Atlanta, GA 30303, USA

E. E. Rutström
Department of Economics, Andrew Young School of Policy Studies, and Dean's Behavioral Economics Laboratory, Robinson College of Business, Georgia State University, Atlanta, GA 30303, USA

G. W. Harrison (✉)
Center for the Economic Analysis of Risk, J. Mack Robinson College of Business, Georgia State University, P.O. BOX 4036, Atlanta, GA 30302-4036, USA
e-mail: gharrison@gsu.edu

Keywords Subjective probabilities · Scoring rules · Risk attitudes

Subjective probabilities are operationally defined as those probabilities that lead an agent to choose some prospects over others when outcomes depend on events that are not yet actualized. These choices could be as natural as placing a bet on a horse race, or as experimentally structured as responding to the payoff prizes provided by some scoring rule. In order to infer subjective probabilities from observed choices of this kind, however, one either has to make some strong assumptions about risk attitudes or jointly estimate risk attitudes and subjective probabilities. We show how the latter can be implemented by pairing several experimental tasks together, some of which identify risk attitudes and some of which identify the interplay between risk attitudes and subjective probabilities. Joint estimation of a structural model of choice across these two types of tasks allows one to make inferences about subjective probabilities from observed behavior in relatively simple choice tasks.

The notion that subjective probabilities can be usefully viewed as prices at which one might trade has been a common one in statistics, and is associated with de Finetti (1937, 1970) and Savage (1971). It is also clear, of course, in the vast literature on gambling, particularly on the setting of odds by bookies and parimutuel markets (Epstein (1977; p. 298ff.)). The central insight is that subjective probabilities of events are marginal rates of substitution between contingent claims, where the contingencies are events that the probabilities refer to. There are then a myriad of ways in which one can operationalize this notion of a marginal rate of substitution.[1]

Scoring rules are procedures that convert a "report" by an individual into a lottery defined over the outcome of some event. The formal link between scoring rules and optimizing decisions by agents is also familiar, particularly in Savage (1971), Kadane and Winkler (1987, 1988), Holt (1986) and Hanson (2003). Jose, Nau, and Winkler (2008) explore the relationship between expected scores, expected utility, and generalized information/entropy measures for several popular scoring rules and the HARA class of utility functions. Alternatives to scoring rules include procedures like those developed for utility elicitation by Becker, DeGroot and Marschak (1964)[2] (BDM) and adapted to eliciting probabilities. Karni (2009), Grether (1992), Köszegi and Rabin (2008), and Holt and Smith (2009) are examples of this alternative.[3] All of these elicitation procedures have some potential problems. The first is the poor incentive properties around the true subjective belief. This is particularly the case for BDM procedures where under-reporting leads to very small expected lost earnings. The second is that explanations of these procedures are not naturally easy to understand for subjects, such that payoff consequences to various decisions must be made transparent.

[1] For example, one could elicit the p that makes the subject indifferent between a lottery paying M with probability p and m with probability (1-p), for M>m, and a lottery paying M if the event occurs and m if it does not (Marschak (1964; p. 107 ff.)). This method formally requires that one elicit indifference, which raises procedural issues that can be avoided by using the type of scoring rules investigated here.

[2] See for example Rutström (1998), Harstad (2000), Plott and Zeiler (2005), and Hao and Hauser (2012) for discussions of properties of BDM.

[3] Let there be two prizes, x>y. The subject reports a probability ξ, and a random number ζ is selected from the unit interval. If $\zeta \leq \xi$ the subject gets the lottery that pays off x if the event occurs, and y otherwise; if $\zeta > \xi$ the subject gets the lottery that pays off x with probability ζ and y with probability 1-ζ.

One advantage with the BDM procedure is that the curvature of the utility function does not confound the elicited probabilities, since the same payoffs can be used in the reference lottery and the actual lottery. Under rank dependent utility (RDU), however, one needs to recognize that it is the decision weights and not the subjective probabilities that are elicited. In the case of scoring rules one could instead undertake "calibrating adjustments" to the elicited beliefs for non-linear utility functions and/or probability weighting, as carried out in Offerman, Sonnemans, van de Kuilen and Wakker (2009; §6).[4] Their elegant approach has a reduced form simplicity, and is agnostic about which structural model of decision making under risk one uses. The maintained assumption, however, is that any deviation in reports from objective and known probabilities transfers directly to a task where probabilities are subjective, less precise and not known with certainty. While precise and agnostic about structural assumptions, the method requires the employment of identical tasks and incentives across the natural lottery for which one wants to elicit subjective beliefs and the calibration task with known objective probabilities. Our approach is more general in the sense that it does not require two identical instruments for eliciting the beliefs and the belief calibration function.

Our approach also has the advantage of allowing a structural identification of sources of imprecision about inferences over subjective probabilities. If inferred subjective probabilities are conditioned on knowing risk attitudes from earlier tasks, or assuming risk attitudes a priori, then any behavioral or statistical uncertainty on the calibration should be allowed to "propagate" into some additional uncertainty over inferences about subjective probabilities. Our joint estimation method allows these error propagation effects to occur, as theory says they should, providing more reliable estimates of subjective probabilities, even if those estimates have large standard errors. Reliability and precision are not the same thing. One can numerically generate a precise estimate based on maintained assumptions that are false, resulting in an unreliable estimator. In other words, it is possible that the choice task for eliciting subjective probabilities generates a point response that appears to be quite precise by itself, but which is actually not a very precise estimate of the latent subjective probability when one properly accounts for uncertainty over the "calibrated" risk attitudes.

We are not the first to propose the general idea of joint estimation of subjective probabilities and utility functions. Viscusi and Evans (1998) introduced the idea in an examination of how subjects appear to update latent subjective probabilities when told how objective probabilities change. They estimate subjective posterior probabilities and utility curvatures from stated willingness to pay for risk reductions in hypothetical surveys, testing whether subjects completely respond to the information given or are influenced by prior beliefs.[5] Viscusi and Evans (2006) extend this approach to compare implied subjective probabilities to stated probabilities also observed in hypothetical

[4] The need for some correction is also recognized by Offerman, Sonnemans and Schram (1996; p.824, fn.8) and Rutström and Wilcox (2009; p.11, fn.8).

[5] Under the assumption that subjects only use information in the survey to form their risk perceptions (Table 2, p. 30, Case 1), they estimate the subjective risk perception. Evans and Viscusi (1993) show that respondents to this survey behave as if they are risk neutral in the face of "small" changes in the risk of using the product. Viscusi and Evans (1990) examine responses to hypothetical surveys of compensating differentials for "large" subjective job risks, using the same general framework, and show that their estimates are invariant to whether they assume linear, logarithmic or exponential utility models, implying that their subjects also behave as if they are risk neutral.

survey responses.[6] We extend the idea of joint estimation to data derived from incentivized popular scoring rules for eliciting subjective probabilities and lottery choices, so that there are real monetary consequences of different reports. Our approach uses the actual incentives provided by these scoring rules, along with explicit structural models of decision-making under risk, to infer the latent subjective probabilities that agents employ.[7]

In Section 1 we briefly state the theory underlying our approach and relate this to the literature on belief elicitation. The properties of the Quadratic Scoring Rule (QSR) and Linear Scoring Rule (LSR), and the fact that responses to these are affected by risk attitudes, are well known. We assume throughout that the agent is acting in what is called a "probabilistically sophisticated" manner, although our method does not restrict the characterization of risk attitudes to expected utility theory (EUT), or to specific functional forms. We also consider the inference of subjective probabilities for subjects who are assumed to make decisions according to the RDU model. This extension is particularly appropriate in the case of eliciting subjective probabilities, because it involves allowing for probability weighting and non-additive decision weights on the utility of final outcomes. Given that one of the probabilities to be weighted is the subjective probability being estimated, one might expect estimates of the subjective probability to be even more sensitive to the correct specification of the model of risk attitudes employed.

In Section 2 we describe the experimental task we posed to 140 subjects, split roughly equally across the QSR and LSR alternatives. Our subjects made choices over a number of standard lotteries, characterized by objective uncertainty over monetary outcomes between $0 and $100. The lotteries vary both in prizes and in probabilities, allowing us to identify parameters for both utility functions and probability weighting functions. Subjects also gave responses to either a QSR or LSR choice task over subjective beliefs. The prizes on each of these scoring rule tasks also spanned $0 and $100, so that we were able to infer risk attitudes over the same prize domain as the scoring rule responses.

Section 3 formally sets out the econometric model used for estimating subjective probabilities, spelling out the manner in which we undertake joint estimation over all tasks in order to identify subjective probabilities. Section 4 then presents our estimates of the inferred subjective probabilities from these scoring rules, after adjusting for risk. Our primary result is that subjective probabilities inferred under the assumption of risk neutrality are very different than the subjective probabilities inferred when one allows the data to say how risk averse the subjects were. This finding has immediate implications for the practical use of scoring rules,[8] which is the focus here; it also

[6] They also differentiate between posterior subjective risk probabilities and the "behavioral probabilities" that agents use when making wage choices conditional on those subjective probabilities. These behavioral probabilities can differ from subjective probabilities, in exactly the same manner as weighted probabilities arise in rank-dependent utility models.

[7] There is nothing in the approach of Viscusi and Evans (1998, 2006) that requires using hypothetical survey data. They do require some identifying assumptions about how agents form their posterior risks, and for that they use a generalized "quasi-Bayesian" learning model with specific distributional assumptions about the data-generating process for risks. It would be valuable to contrast, in a controlled manner with experiments using real incentives, the effect of these different approaches to identifying utility functions and subjective probabilities.

[8] There is a large, practical literature on the "normative" elicitation of subjective probabilities, reviewed by O'Hagen et al. (2006) and illustrated well by Shephard and Kirkwood (1994). One characteristic of those elicitation procedures is that they involve considerable real-time, one-on-one feedback between the elicitor and the elicitee, often including results from proper scoring rules. Our approach is to better model inferences from more static, impersonal applications of those scoring rules, as a prelude to the design and evaluation of normative elicitation procedures with incentives.

has implications for inferences that can be drawn from prediction markets since experimental findings consistently show that people are risk averse on average. We offer a methodology for jointly estimating risk attitudes and subjective probabilities, and show that utility curvature and probability weighting can have opposing qualitative effects on inferences about subjective probabilities. Section 5 draws conclusions.

1 Scoring rules

For simplicity we assume throughout that the events in question only have two outcomes.[9] A scoring rule asks the subject to make some report θ, and then defines how an elicitor pays a subject depending on their report and the outcome of the event. This framework for eliciting subjective probabilities can be formally viewed from the perspective of a trading game between two agents: you give me a report, and I agree to pay you $X if one outcome occurs and $Y if the other outcome occurs. The scoring rule defines the terms of the exchange quantitatively, explaining how the elicitor converts the report from the subject into a lottery. We use the terminology "report" because we want to view this formally as a mechanism, and do not want to presume that the report is in fact the subjective probability π of the subject. In general, it is not.

The QSR was apparently first used by McKelvey and Page (1990), and later by Offerman, Sonnemans and Schram (1996), McDaniel and Rutström (2001), Nyarko and Schotter (2002), Schotter and Sopher (2003), Costa-Gomes and Weizsäcker (2008) and Rutström and Wilcox (2009).[10] In each case the subject is implicitly or explicitly assumed to be risk-neutral.[11] Scoring rules that are linear in the absolute deviation of the estimate have been used by Dufwenberg and Gneezy (2000) and Haruvy, Lahav and Noussair (2007). Croson (2000) and Hurley and Shogren (2005) used scoring rules that are linear in the absolute deviation as well as providing a bonus for an exactly correct prediction. Scoring rules that provide a positive reward for an "exact" prediction and zero otherwise have been used by Charness and Dufwenberg (2006).

In many cases the inferential objective has been to test hypotheses drawn from "psychological game theory," which rest entirely on making operational the beliefs of players in strategic games. For this purpose, and also in other applications, the elicitation of beliefs is combined with some other experimental task. Other applications include testing the hypothesis that the belief elicitation task will encourage players in a game to think more strategically (Croson (2000), Costa-Gomes and Weizsäcker (2008), Rutström and Wilcox (2009)). Of course, combining tasks in this way violates the "no stakes condition" required for the QSR to elicit beliefs reliably unless one assumes that the subject is risk neutral (Kadane and Winkler (1988), Karni and Safra (1995)

[9] Extensions to eliciting subjective beliefs over continuous events are considered by Matheson and Winkler (1976) and Harrison, Martínez-Correa, Swarthout and Ulm (2012).

[10] Hanson (1996) contains some important corrections to some of the claims about QSR elicitation in McKelvey and Page (1990).

[11] McKelvey and Page (1990) augmented the scoring rule procedure with a "binary lottery" payment procedure to induce risk-neutrality. In theory the subject earns "points" in the scoring rule, which convert in a linear manner into an increased probability of winning some lottery defined over a high prize and a low prize. There is considerable controversy over the behavioral validity of this procedure, reviewed in Harrison, Martínez-Correa and Swarthout (2013a, b).

and Karni (1999)). Only one of these studies employs a "spectator" treatment in which players are asked to provide beliefs but do not take part in the constituent game determining the event outcome: study #2 of Offerman, Sonnemans and Schram (1996).

The popular QSR is defined in terms of two positive parameters, α and β that determine a fixed reward the subject gets and a penalty for error. Assume that the possible outcomes are A or B, where B is the complement of A, that θ is the reported probability for A, and that Θ is the true binary-valued outcome for A. Hence $\Theta=1$ if A occurs, and $\Theta=0$ if it does not occur (and thus B occurs instead). The subject is paid $S(\theta|\text{A occurs})=\alpha-\beta(\Theta-\theta)^2=\alpha-\beta(1-\theta)^2$ if event A occurs and $S(\theta|\text{B occurs})=\alpha-\beta(\Theta-\theta)^2=\alpha-\beta(0-\theta)^2$ if B occurs. In effect, the score or payment penalizes the subject by the squared deviation of the report from the true binary-valued outcome, Θ, which is 1 and 0 respectively for A and B occurring. An omniscient seer would obviously set $\theta=\Theta$. The fixed reward is a convenience to ensure that subjects are willing to play this trading game, and the penalty function simply accentuates the penalty from not being an omniscient seer. In our experiments $\alpha=\beta=\$100$, so subjects could earn up to \$100 or as little as \$0. If they reported 1 they earned \$100 if event A occurred or \$0 if event B occurred; if they reported ¾ they earned \$93.75 or \$43.75; and if they reported ½ they earned \$75 no matter what event occurred.

It is intuitively obvious, and also well known in the literature (e.g., Winkler and Murphy (1970) and Kadane and Winkler (1988)), that risk attitudes will affect the incentive to report one's subjective probability "truthfully" in the QSR.[12] A sufficiently risk averse agent is clearly going to be drawn to a report of ½, and varying degrees of risk aversion will cause varying distortions in reports from subjective probabilities. If we knew the form of the (well-behaved) utility function of the subjects, and their degree of risk aversion, we could infer back from any report what subjective probability they must have had. Indeed, this is exactly what we do below, recognizing that we only ever have estimates of their true degree of risk aversion.

The LSR is also defined in terms of two positive parameters, γ and λ, that serve as fixed rewards and penalties, respectively. The subject is paid a fixed reward less some multiple of the absolute difference between their report and what actually happened, which is also what an omniscient seer would have reported. Thus the payment is $S(\theta|\text{A occurs})=\gamma-\lambda(1-\theta)$ if event A occurs and $S(\theta|\text{B occurs})=\gamma-\lambda(\theta-0)$ if B occurs. We again set $\gamma=\lambda=\$100$, generating payoffs of \$100 or \$0 for a report of 1; \$75 and \$25 for a report of ¾; and \$50 no matter what the outcome for a report of ½. The LSR is not a favorite of decision theorists, since a risk neutral subject would jump to corner-solution reports of 1 or 0 whenever their true beliefs were either side of ½. But when the subject is (even modestly) risk averse, an interior solution is obtained, and we face the same issues of inference as with the QSR. The LSR is a favorite of experimental economists because of the simplicity of explaining the rule: the score for a report and an

[12] There exist mechanisms that will elicit subjective probabilities without requiring that one correct for risk attitudes, such as the procedures proposed by Köszegi and Rabin (2008; p.199), Karni (2009), Grether (1992), Holt and Smith (2009), Offerman, Sonnemans, van de Kuilen and Wakker (2009) and Hao and Houser (2012), discussed further below. The last four employ these mechanisms in an experimental evaluation. We discussed earlier the difficulties of practical application of these methods.

event is linear in the reported probability report, so there is no need for elaborate tables showing cryptic payoff scores for discrete reports.[13]

In order to avoid portfolio effects from the combined choices rewards in these scoring rule task are sometimes very, very small. For example, Nyarko and Schotter (2002) and Rutström and Wilcox (2009) gave each subject an endowment of 10 cents, from which their penalties are to be deducted. In the latter study this does not present a problem since the focus is not on analyzing the elicited beliefs but on analyzing the game play as a function of exposing subjects to a belief elicitation process. Nevertheless, one has to worry about the incentive properties of the elicitation method once the interest is on analyzing the beliefs themselves.

The need to calibrate or control for risk aversion is often not made explicit, or is claimed to be of marginal concern. Schotter and Sopher (2003; p. 504) recognize the role of risk aversion, but appear to argue that it is not a factor behaviorally:

> It can easily be demonstrated that this reward function provides an incentive for subjects to reveal their true beliefs about the actions of their opponents. Telling the truth is optimal; however, this is true only if the subjects are risk neutral. Risk aversion can lead subjects to make a "secure" prediction and place a .50 probability of each strategy. We see no evidence of this type of behavior.

Of course, evidence of subjects selecting the probability report of ½ only shows that the subject has extreme risk aversion. The absence of that extreme evidence says nothing about the role that risk aversion might play in general. Only one QSR study attempts to explicitly calibrate the beliefs for non-linear utility functions and/or probability weighting: Offerman, Sonnemans, van de Kuilen and Wakker (2009; §6). We introduce an alternative experimental method where the decision model can be identified econometrically from one set of tasks for the same subjects, or even on a different pool of subjects drawn from the same population, and then statistically integrated with the belief elicitation task, while transparently allowing error terms to propagate. The estimation of the decision model over risk and subjective probabilities is joint and simultaneous, even if one can think of the lottery choices as recursively identifying the decision model.

2 Experimental design

Figure 1 illustrates the lottery choice that our subjects were given. Each subject faced 45 such choices, where prizes spanned the domain $0 up to $100 and probabilities for various prizes varied across each lottery. One choice was selected to be paid out at random after all choices had been entered. Choices of indifference were resolved by rolling a die and picking one lottery, as had been explained to subject. This interface

[13] Hanson (1996; p. 1224) provides a useful reminder that discrete implementations of proper scoring rules can also engender piecewise linear opportunity sets. He points out that certain regions of the QSR implemented by McKelvey and Page (1990) were actually LSR, and that risk-neutral subjects would then rationally report a probability at the extremes of that linear region, and not at the discrete alternative closest to their true belief.

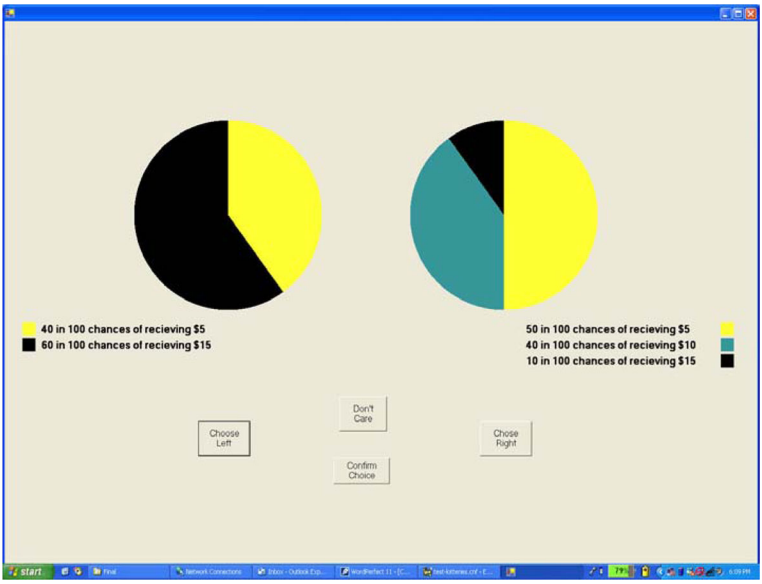

Fig. 1 Illustrative lottery choice

builds on the classic binary choice design of Hey and Orme (1994), and is discussed in greater detail in Harrison and Rutström (2008; Appendix B). The lotteries were presented sequentially in 3 blocks of 15, where each block had prizes in one of three intervals between $0 and some higher level. One level was between $0 and $1, the other level was between $0 and $10, and the third level was between $0 and $100. We presented the lotteries sequentially so that the subject could see that all of the lotteries in one block were for a given scale. The sequence of blocks was randomized across subjects. Complete instructions are provided in online Appendix A and the full set of lotteries in online Appendix D.

The belief tasks were presented to subjects with a novel interface that has many attractive features.[14] Figure 2 shows the interface for the QSR as it was presented to subjects on a computer screen and in printed instructions. The interface was explained with instructions which used a trusty old bingo cage to illustrate one underlying random process. By varying the slider the subject could choose a report, with the conditional payoffs being displayed. The formula for the scoring rule was not shown, but the subject instead just saw alternative payoffs for different reports. The subject was then taken through displays of their payoffs if they chose to report 0% or 100%.[15] Each subject participated in an unpaid training choice, in which they were told the number of orange balls in the bingo cage that was on public display, and asked to make a report

[14] This interface immediately extends to other tasks with a cardinal scale that experimental economists use to elicit risk preferences, discount rates, or social preferences. We opted to use binary choices for our lottery tasks, to be consistent with the vast bulk of the literature on the elicitation of risk preferences.

[15] The display of the probability on the right side of the slider always shows an integer percentage, and the earnings were always calculated on that value, thus making the continuous probability scale into an integer one. In the statistical analysis we treat the probability as taking on 101 integer values in the range 0 to 100.

and confirm it. We deliberately adopted an extremely high scale of a maximum $1,000 payoff to ensure that the subjects understood that this was to be a trainer.

Each subject then participated in 7 belief elicitation tasks, knowing that one would be selected for payment. The first 3 were repetitions of the training task with orange and white ping pong balls, but with completely different distributions of orange and white ping pong balls, and served to provide subjects with hands-on, incentivized experience in the scoring rule.[16] We do not analyse these choices here. The fourth task was based on the outcomes of a test in psychology for empathy known as the Eyes Test (e.g., Baron-Cohen (2003). All subjects had completed this test at the outset of the session, and the event they were asked about was whether the score that a randomly chosen man got on the Eyes Test was equal to or greater than the score that a randomly chosen woman would get. The final three tasks were based on the 2008 U.S. presidential election, which was to be held about 1 week after the session. One task was whether the outright winner of the presidency would be a Democrat or a Republican, one task was whether or not the winning share of the popular vote would be at least 5 percentage points greater than the losing share, and the final task was whether or not the winning share of the popular vote would be at least 10 percentage points greater than the losing share.[17] Our own a priori expectations for these subjective probabilities were just around 50%, around 80%, around 65% and less than 10%, respectively.

The events were explained in written instructions, which were also read out loud. The event based on the Eyes Test was explained as follows: "We will pick one man and one woman in the room. Do you think the man who is selected will have a higher score on the Eyes Test than the woman who is selected?" We then explained how we would randomly select one man and one woman in the session and compare their scores. They were then asked to bet on: "That the man we select at random will have a higher score on the Eyes Test than the woman we select at random."

After subjects had completed their bets on the bingo cage and Eyes Test tasks, the final three events, about the presidential elections, included the following three bets:

1. Will the next President of the United States be a Democrat?
2. Will the popular vote for the winning candidate be 5 or more percentage points greater than the popular vote for the losing candidate?
3. Will the popular vote for the winning candidate be 10 or more percentage points greater than the popular vote for the losing candidate?

We explained that the first question was about the outcome of the Electoral College vote, and not the popular vote, and that for the second and third question, we were asking if they thought that the winner of the popular vote would beat the loser by 5 or 10

[16] The subjects were told that there were 60 balls in total in a publicly visible, but initially covered, bingo cage, but were not told the number of orange or white balls. The urn was uncovered and spun for 10 rotations, and then the subject had to make a report that a ball drawn at random would be orange. Applying the methods examined here to these data does not change any of our conclusions.

[17] These events compare to similar events employed in popular prediction markets, inspired by Forsythe, Nelson, Neumann and Wright (1992). See http://www.biz.uiowa.edu/iem/ for the current version of this market, and the contracts traded in the 2008 presidential election.

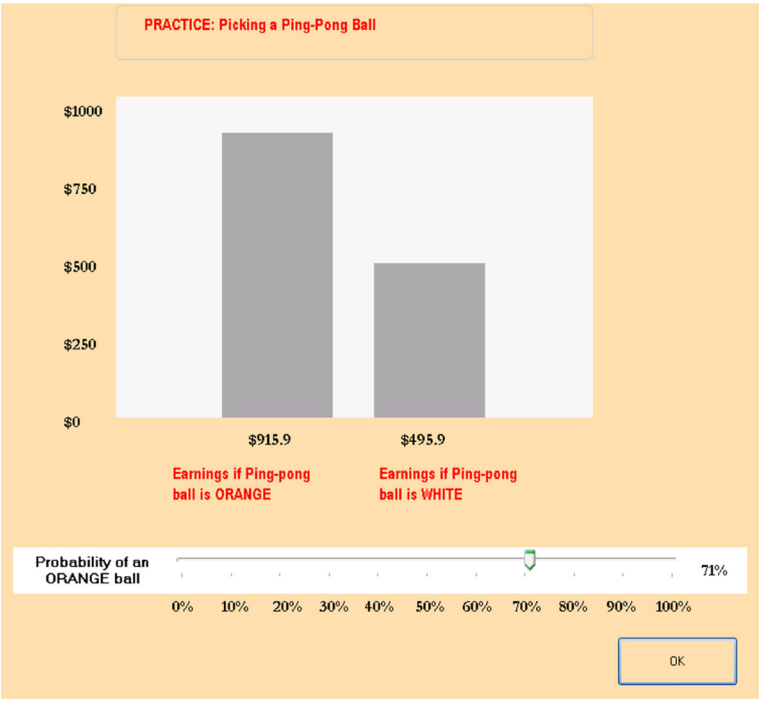

Fig. 2 Illustrative quadratic scoring rule interface

percentage points or more. The subjects then completed their belief elicitation tasks for these presidential election events, and went on to the lottery choice tasks described earlier.

We recruited 140 subjects from the student population of the University of Central Florida, split equally across the QSR and LSR treatments. The experiments were conducted in the week prior to the 2008 election, Monday October 27 through Friday October 31. Our 140 subjects earned about $90 on average for the two paid tasks. Each session lasted around 1½ to 2 hour, and never more than 2 hour. There was considerable variation in earnings, with one subject taking home $3 and another subject taking home $205.

3 Econometric model

We develop the econometric model to be estimated in three stages, with details presented in online Appendix C. First we specify risk attitudes assuming an EUT model of latent choice, where the focus is entirely on the concavity of the estimated utility function. Second, we specify risk attitudes assuming a RDU model of latent choice, so that risk attitudes are determined by the interplay of concave utility functions and non-linear probability weighting.[18] Third, we consider the joint estimation of risk attitudes and subjective probability, using either the EUT or the RDU specification.

[18] We could just develop a RDU model and test if the estimated probability weighting is the identity function, in which case the RDU model collapses to an EUT model. However, the exposition is, in our view, simpler if one develops the models separately because of the familiarity of EUT to most economists.

3.1 Estimating the EUT and RDU models

We assume an Expo-Power (EP) utility function originally proposed by Saha (1993). Following Holt and Laury (2002), the EP function is defined as

$$u(y) = \left[1-\exp\left(-\alpha y^{1-r}\right)\right]/\alpha, \quad (1)$$

where α and r are parameters to be estimated, and y is income from the experimental choice. The EP function can exhibit increasing or decreasing relative risk aversion (RRA), depending on the parameter α: RRA is defined by $r+\alpha(1\text{-}r)y^{1\text{-}r}$, so RRA varies with income if $\alpha \neq 0$ and the estimate of r defines RRA at a zero income. This function nests CRRA (as $\alpha \rightarrow 0$) and CARA (as $r \rightarrow 0$). The functional form of utility employed here is of no importance, and any monotonic increasing function of u(.) could have been implemented. We can identify both of these parameters because of the variation in prizes that are included in the full lottery design, as shown in online Appendix D.

The RDU model extends the EUT model by allowing for decision weights on lottery outcomes. We use the same utility function specification (1), but allow for decision weights generated by a probability weighting function. We adopt the simple "power" probability weighting function proposed by Quiggin (1982), with curvature parameter γ:

$$\omega(p) = p^{\gamma} \quad (2)$$

So $\gamma \neq 1$ is consistent with a deviation from the conventional EUT representation.[19] Any other functional form could have been used, as long as the probability function maps the probability $p=[0,1]$ into the unit interval [0,1] and the function is strictly increasing in p.

Figure 3 shows the manner in which the parameter γ characterizes the probability weighting function and the decision weights used to evaluate lottery choices, when the outcomes are ranked from best to worst. Since we assume $\gamma=0.77<1$ in this illustration, to anticipate our estimates, the probability weighting function $\omega(p)$ is concave in p. For simplicity here we assume lotteries with 2, 3 or 4 prizes that are equally likely when we generate the decision weights. So for the case of 2 prizes, each prize has $p=½$; with 3 prizes, each prize has $p=⅓$; and with 4 prizes, each prize has $p=¼$. We see the usual result, that the decision weights on the largest prizes are relatively greater than the true probability, and the decision weights on the smallest prizes are relatively smaller than the true probability, reflecting optimism over the outcomes.[20]

Each panel in Fig. 3 is important for our analysis. For the purposes of estimating γ from the observed lottery choices with known probabilities we only need the decision weights in the right panel of Fig. 3. But for the purposes of recovering a subjective probability π subject to probability weighting, as distinct from p, the induced objective probability, we instead only need the probability weighting function. In fact, we need

[19] We compared our results to estimations based on the inverse-S shaped funtion, popularized by Tversky and Kahneman (1992) and the flexible Prelec (1998) function that allows a variety of shapes including concave, convex, S-shaped and inverse S-shaped. We conclude that the power function fits better for these data. It has a better log-likelihood, and it rejects a linear probability weighting function. The shape is confirmed in our estimates of the flexible Prelec function, while the inverse-S function is close to linear, which is the best it can do to proxy a power function.

[20] Hence a concave probability weighting function, as in the left panel of Figure 3, implies risk-seeking behavior, *ceteris paribus* the curvature of the utility function.

its inverse function, since it is the π in the $\omega(\pi)$ function that we are seeking to recover in that case. We do not directly observe $\omega(p)$ or $\omega(\pi)$, but we can estimate $\omega(\cdot)$ as part of the latent structure generating the observed choices in the two types of task, implicitly assuming that $\omega(p)=\omega(\pi)$. To anticipate slightly the exposition below, once we have $\omega(\cdot)$ we can then recover π by directly applying the estimated probability weighting function, such as the one shown, for a typical γ, in the left panel of Fig. 3.[21]

3.2 Estimating the subjective probability

To estimate the subjective probability π that each subject holds from LSR or QSR responses we have to assume something about how they make decisions under risk. This is obvious in theory, and the only issue then is how to operationalize that property of these scoring rules.

If they are assumed to be risk neutral, then we can directly infer the subjective probability from the report of the subject.[22] This result is immediate under the QSR, but raises a problem of interpretation under the LSR if the reports are not at the corner solutions of 0% and 100%. On the other hand, any minimal level of risk aversion will suffice, under the LSR, to generate interior responses, so we assume that the subjects indeed have some minimal level of risk aversion when we report "risk neutral subjective beliefs" for the LSR.

Moving to the models that allow for varying risk attitudes, we jointly estimate the subjective probability and the parameters of the core model. Assume for the moment that we have an EUT specification. The subject who selects report θ from a given scoring rule receives the following EU

$$\mathrm{EU}_\theta = \pi_\mathrm{A} \times \mathrm{u}\Big(\text{payout if A occurs}\Big|\text{report } \theta\Big) + (1-\pi_\mathrm{A}) \times \mathrm{u}\Big(\text{payout if B occurs}\Big|\text{report } \theta\Big) \tag{3}$$

where π_A is the subjective probability that A will occur. The payouts that enter the utility function are defined by the scoring rule and of course the specific report θ, and span the interval [\$0, \$100]. For the QSR and a report of 75%, for example, we have

$$\mathrm{EU}_{75\%} = \pi_\mathrm{A} \times \mathrm{u}(\$93.75) + (1-\pi_\mathrm{A}) \times \mathrm{u}(\$43.75) \tag{3'}$$

For the LSR, and the same report, we have:

$$\mathrm{EU}_{75\%} = \pi_\mathrm{A} \times \mathrm{u}(\$75) + (1-\pi_\mathrm{A}) \times \mathrm{u}(\$25) \tag{3''}$$

[21] There is a caveat to this intuition when subjective probabilities are close to 0.5, due to a jump discontinuity in decision weights as reports under the QSR vary around 0.5. At the point where the report is 0.5 the rank ordering of the choice options reverses. For reports less than 0.5 the A option implies higher monetary rewards, but for reports greater than 0.5 it is the B option that offers the higher reward. For sufficiently large probability weighting this jump discontinuity can wreak havoc with the ability to infer latent subjective probabilities. In our applications the probability weighting is not severe, and we can initialize the maximum likelihood estimation at the solution values obtained by assuming EUT (so that directional derivatives do not need to be numerically evaluated in the problematic regions of γ and π).

[22] The expression "risk neutral" here should be understood to include the curvature of the utility function and the curvature of the probability weighting function. So it is not just a statement about the former, unless one assumes EUT.

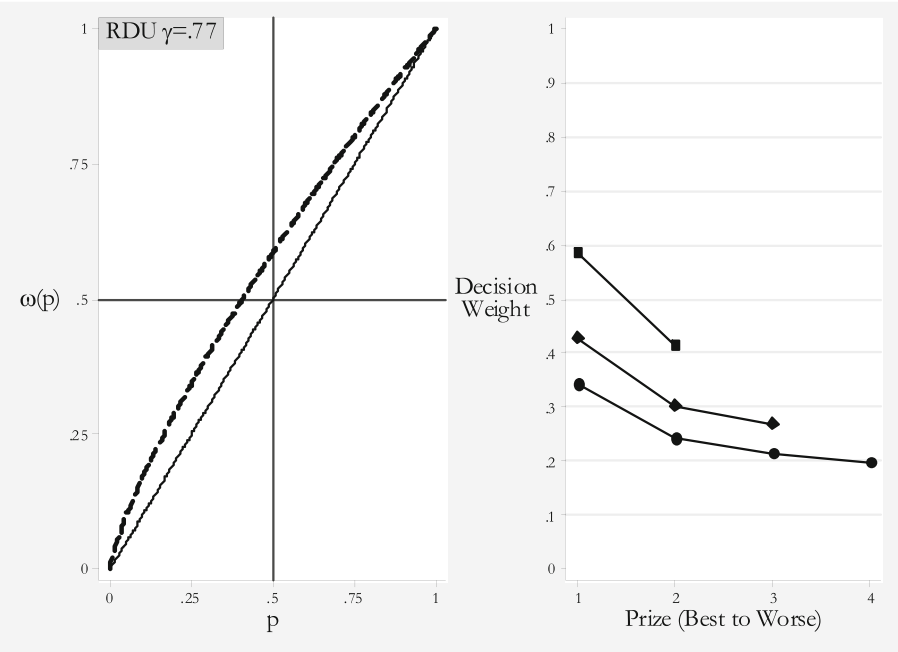

Fig. 3 Probability weighting and decision weights

and so on for other possible reports. We observe the report made by the subject for QSR or LSR. This report can take 101 different integer values defined over percentage points. Then we can calculate the likelihood of that choice given values of r, π_A and μ, where the likelihood is the multinomial analogue of the binary logit specification used for lottery choices. We define

$$eu_{\Theta} = \exp[(EU_{\Theta}/v)/\mu] \tag{4}$$

for any report Θ, where μ is a Fechner error and v is a contextual utility transformation, both discussed in online Appendix C, and then

$$\nabla EU = eu_{\theta}/(eu_{0\%} + eu_{1\%} + \ldots + eu_{100\%}) \tag{5}$$

for the specific report θ observed, analogously to the comparable expression for EUT or RDU

We need r and α to evaluate the utility function in (3), we need π_A to calculate the EU_{θ} in (3) for each possible report Θ in {0%, 1%, 2%, …, 100%} once we know the utility values, and we need μ to calculate the latent indices (4) and (5) that generate the subjective probability of observing the choice of specific report θ when we allow for some noise in that process. The joint maximum likelihood problem is to find the values of these four parameters that best explain observed choices in the belief elicitation tasks as well as in the lottery tasks. We estimate these parameters simultaneously, even if the lottery tasks could be used solely to estimate the r and α parameters, and are needed to ensure identification of all parameters.

Exactly the same logic extends to the model in which we assume an RDU latent structure instead of an EUT latent structure. In effect, the lottery task allows us to identify r

and α under EUT, and r, α *and* γ under RDU, thanks to the variations in both prizes and probabilities in this task. Individual heterogeneity is allowed for by estimating both risk attitudes and subjective probabilities as linear functions of the demographic characteristics defined earlier.

4 Results

4.1 Raw elicited beliefs

Figure 4 displays the raw responses from each of the scoring rules for each event, in the form of kernel densities, and Table 1 shows summary statistics of the elicited responses. The four events are the "Eyes Test," the "President," "Win by 5%," and "Win by 10%." The summary statistics suggest that the QSR and LSR provided roughly the same responses, but the densities in Fig. 4 do have some differences in shape. In part this simply alerts us to be aware of the non-Gaussian shape of these distributions.

The general location of the densities corresponds with our qualitative priors on the subjective beliefs that were to be expected for these events. Recall that the Eyes Test asked for the probability that a typical male would score better than a typical female. For the Eyes Test it appears, from the observation that the modal response is around 0.4 for both LSR and QSR treatments, that the sample did not expect the male score to exceed the female score, but that there was a wide variability around this modal belief. The sample appeared confident that Barack Obama would indeed win the election outright, but displayed a healthy sense of perspective on what the winning margin

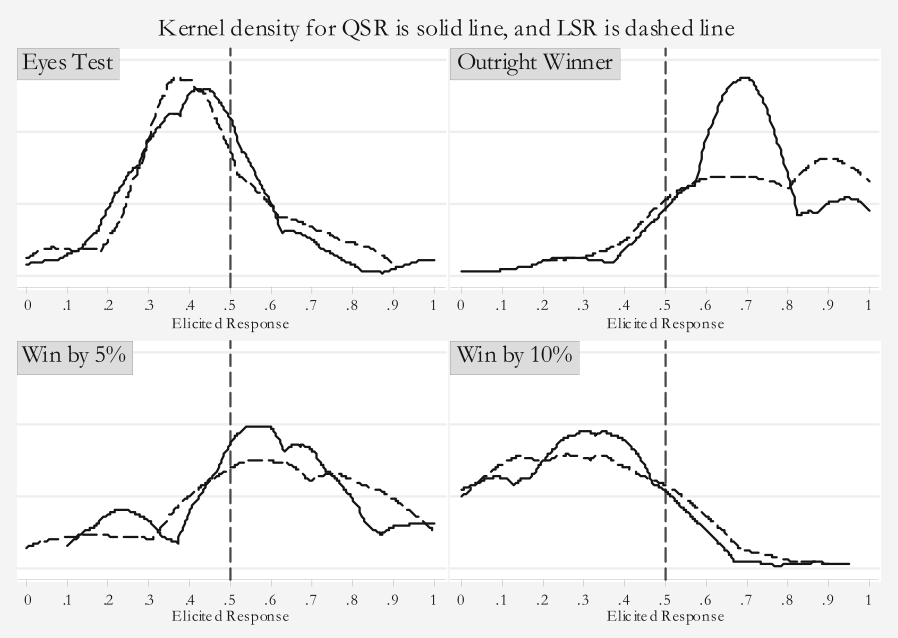

Fig. 4 Elicited responses from QSR and LSR

Table 1 Descriptive statistics for scoring rule responses

Event	Scoring rule	Mean	Median	Standard deviation
Eyes Test	Quadratic	0.43	0.4	0.19
	Linear	0.43	0.4	0.19
	Both	0.43	0.4	0.19
President	Quadratic	0.69	0.7	0.2
	Linear	0.74	0.75	0.22
	Both	0.71	0.7	0.21
Win by 5%	Quadratic	0.59	0.6	0.23
	Linear	0.58	0.6	0.26
	Both	0.59	0.6	0.24
Win by 10%	Quadratic	0.28	0.3	0.2
	Linear	0.29	0.26	0.21
	Both	0.28	0.3	0.21

would be. For our purposes, the choices of a 5% and 10% threshold for the popular vote could not have worked out better, with a majority believing that a 5% margin would be attained but that a 10% margin would not. These results show, at a minimum, that responses were at least correlated with what we believe to be reasonably coherent subjective beliefs for these events.

The fact that the responses to the LSR are not at "corner" values of 0 or 1 shows that the subjects were not exactly risk neutral. But it does not show much more, because one would observe some interior response even for small amounts of risk aversion, as noted earlier.

With the exception of the bets on the outright winner of the presidential election, the distribution of responses for the two scoring rules are roughly the same. This conclusion is supported by a Kolmogorov-Smirnov test of the null hypothesis that the two distributions are equal. The exception is the case of the outright winner event, where the p-value is only 0.011, so we can reject that hypothesis in this instance. This finding provides some support for those that would prefer to use the LSR on the grounds that it is simpler to explain to subjects than the QSR. Of course, the real issue is whether they generate the same estimates of subjective probability when one allows for risk attitudes.

4.2 Characterizing risk attitudes

Looking just at the lottery choices under a maintained hypothesis of EUT for now, we find evidence of modest risk aversion at low stakes (since $r=0.3>0$, and r defines RRA at $y=0$), and evidence of slightly increasing relative risk aversion as the prizes climb to \$100 (since $\alpha=0.03>0$). Detailed results are provided in online Appendix B, since they are only of indirect interest here. Given these parameter estimates we can calculate RRA at various prize levels: at \$25, \$50, \$75 and \$100 the RRA is estimated to be 0.49, 0.61, 0.71 and 0.81, respectively.[23] Thus, despite the relatively low estimate of α the implied risk aversion at higher stakes is much stronger than for the lower stakes.

[23] We simply use the estimated model to predict the point estimate of the RRA for each subject and each of these prizes, and report the average of those RRA point estimates here.

When we allow for an array of covariates to better characterize the heterogeneity of risk attitudes, we observe females to be significantly more risk averse, with RRA 0.14 higher at the $0 level. As expected, allowing for covariates has no significant average effect on the RRA by prize level.

These results suggest that one might see somewhat different effects of "risk-conditioning" in the reports for scoring rules depending on the stakes involved. Our stakes are large in relation to the literature: a maximum prize of $100, compared to common implementations in experiments of a maximum prize of less than $1.[24] In applications where the interest is directly on eliciting beliefs, such small stakes will not generally result in precise estimates, so larger stakes are necessary. Our results on risk attitudes for low stakes and high stakes imply that the extent of adjustment for risk attitudes is much greater for higher stake elicitations. It is a factor for both, since we estimate RRA to be positive for the lowest stakes, but it is not as serious a factor as when the stakes are lower.

The results from estimating the RDU model are slightly different. Detailed estimates are again reported in online Appendix B. The estimates of the utility function parameters ŕ and ά are both larger than their counterparts under EUT, implying greater concavity of the utility function. We estimate the probability weighting parameter under RDU to be γ=0.72 without covariates, and can reject the hypothesis that this is equal to 1 (p-value <0.001). A likelihood ratio test of the hypothesis that the EUT model and the RDU model are the same when there are no covariates has a χ^2_1=35.47 (p-value <0.01), so we reject that null. The same conclusion is true when we account for covariates and heterogeneity of responses. There are no strikingly different coefficient estimates due to demographic characteristics (even female is not significant here), and the small changes across the board do not add up to a statistically significant difference. A likelihood ratio test of the hypothesis that the EUT model and the RDU model are the same when including covariates has a χ^2_{14}= 47.71 (p-value <0.01). Thus, we find support for the hypothesis of probability weighting in this case, although here we are not investigating to what extent this is generally true or just holds for portions of our subjects and/or tasks. Since the estimated γ is less than 1, we conclude that subjects are optimistic (or, which is the same, risk seeking), which explains why the utility function is more concave than for EUT when estimated on the same choice data.

One noteworthy feature of these estimates is that one can reject the CRRA specification for both EUT and RDU models in this case. Of course, one might accept CRRA if estimating risk attitudes over a much smaller income domain, such as between $0 and $10, or when the variation in stakes as a percentage is relatively small.

4.3 Estimating subjective probabilities

Table 2 lists the main results from estimating subjective probabilities for each of the four events considered here, and assuming either an EUT or RDU specification. We

[24] To be fair, those low-stake implementations are often in the context of the probability elicitation task being paired with another task, such as the choice of a strategy in a game, and the stake for the probability elicitation task is kept small to avoid the subject attempting to construct a portfolio of paired responses across the two tasks.

Table 2 Estimated subjective probabilities

Event	Specification (log-likelihood)	Point estimate	Standard error or standard deviation	95% confidence interval
Eyes Test	EUT (−4563.2)	0.42	0.022	0.38/0.46
	RDU (−4550.6)	0.31	0.037	0.24/0.39
	Raw responses	0.43	0.19	0/0.95
President	EUT (−4558.1)	0.8	0.033	0.74/0.87
	RDU (−4542.1)	0.87	0.079	0.72/1
	Raw responses	0.71	0.21	0.20/1
Win by 5%	EUT (−4593.8)	0.62	0.030	0.56/0.68
	RDU (−4580.3)	0.57	0.048	0.47/0.67
	Raw responses	0.59	0.24	0.10/1
Win by 10%	EUT (−4560.9)	0.18	0.038	0.11/0.25
	RDU (−4539.5)	0.004	0.024	0/0.05
	Raw responses	0.28	0.2	0/0.77

pool data over LSR and QSR, and control for the effect of the scoring rule with a binary dummy variable.

Assume the EUT specification for now. Given that we find evidence of risk aversion in our subjects over the domain of prizes used in the belief elicitation tasks, our estimated subjective probabilities are all translations of the raw responses away from the 50% response. In both the LSR and QSR we expect risk averse subjects to make choices biased towards 50%, so that when we correct for their risk attitudes the inferred probabilities should move away from 50%. The reason, again, is that risk averse subjects are drawn to respond toward 50% simply to reduce the uncertainty over payoffs, so evidence of risk aversion implies that their true, latent, subjective probabilities must be further away from 50% than their raw responses. Our maximum likelihood estimates simply impose some parametric structure on that theoretical structure, to be able to quantify the extent of the required translation and the precision of the resulting inference about the latent subjective probability.

For the RDU specification, where we find a stronger concavity of the utility function and therefore a stronger effect moving the report closer to the 50% point, there is also an opposite effect moving the report away from the 50% point due to the concavity of the probability function. Our results demonstrate that the net effect of these two forces varies with the closeness to the 50% point. The closer we are to a true, latent subjective probability of 50% the smaller is the utility correction effect but the stronger is the decision weight correction effect.[25]

[25] The magnitude of the probability weighting correction close to 50% is a function of the use of a power function. We also estimated the inverse-S probability weighting function proposed by Tversky and Kahneman (1992) as well as a flexible Prelec (1998) function. The inverse-S function has both concave and convex portions and our estimates move this function as close to the everywhere-concave range as is possible, but with a much lower log likelihood than the power function. The flexible Prelec function is estimated to be concave. We therefore report results only for the power function.

For the Eyes Test, we observe a very small movement in subjective probabilities away from the raw responses under the assumption of EUT. For any given utility curvature, true subjective probabilities that are closer to 50% exhibit a smaller absolute bias from the tendency of aversion to outcome variability to move reports towards 50%. The effect of probability weighting in the RDU model is statistically significant, and the net effect of probability weighting and utility curvature is to further adjust the inferred probabilities away from the 50% responses. However, the net effect masks the fact that there are two underlying forces moving in opposite direction. Correcting for probability weighting will increase the estimated probability, while correcting for utility curvature will decrease the estimated probability. The reason that the estimated probability increases when correcting for probability weighting in this case is because the report we ask subjects to make is for what they perceive as the unlikely event, with a latent subjective probability less than 0.5. The scoring rule pays a lower prize for the unlikely event, and with optimistic probability weighting the worse outcome has a decision weight that understates the subjective belief. The estimated value of the probability weighting parameter for the Eyes Test is γ=0.77, which is the value used to generate the illustration in Fig. 3. Notice also that the probability weighting effect on the report is larger the closer we are to reports of 50%. The strong decrease in the latent probability due to utility curvature is because the estimated utility curvature is so much stronger under RDU than under EUT. The 95% confidence interval does not include the raw report for this test under RDU, but does under EUT.

For the three election events, we see more interesting effects of adjusting for risk aversion. In the "Win by 5%" event, which is again one in which the raw responses are relatively close to 50%, we again infer a small translation from the raw response average of 59% to 62% under EUT, and from 59% to 57% under RDU. Because we are now on the opposite side of the 50% divide, the utility correction will increase the latent probability, while the decision weight correction will cause it to decrease. The other two election events illustrate when one might expect to see the effect of utility curvature exert a more significant quantitative effect because we are further away from the 50% point, and the spread in the payoffs is larger. On the other hand, we expect a smaller effect from decision weight corrections since we are closer to the tails of the cumulative distributions shown in Fig. 3, and hence closer to the fixed points $\omega(\pi)=\pi$ when π is equal to 0 or 1. In the case of the outright winner event, labeled "President" in Table 2, we estimate latent subjective probabilities to be 80% (EUT) or 87% (RDU), rather than the raw response average of 71%. Moreover, the 95% confidence interval on these estimates does not include the raw response. Similarly, in the case of the chance of the popular vote for the winner being more than 10% of the popular vote of the loser, we estimate subjective probabilities of 18% (EUT) or 0.4% (RDU), rather than the raw response average of 28%.

It is possible to estimate the structural model of risk attitudes and subjective beliefs allowing for observable covariates on each parameter. To illustrate, Table B3 in online Appendix B lists details of EUT estimates for the model of the "Win by 5%" event

when we include covariates.[26] One reason for controlling for covariates is to allow for sample composition differences between treatments.[27] We observe no statistically significant effect from using the QSR or LSR for this event, but there is no reason to expect one after we properly condition for risk attitudes. That is, the two scoring rules simply provide subjects with different lotteries with which to place bets about their subjective beliefs. So the same subjective belief should be estimated from each treatment, once one has conditioned on risk attitudes. We find a significant difference between QSR and LSR only for the event "Win by 10%", which generates higher probability estimates in the LSR.

To put these results into perspective, particularly those for the outright winner, it is important to note that in the week of the experiments the tide of public opinion clearly favored Barack Obama to win in a landslide. Eliciting a raw belief of only a 70% chance of Obama winning is therefore puzzling, and an a priori challenge for those that would use the raw results from a QSR or LSR procedure. For example, consider the Iowa Presidential Market, the current version of the prediction market developed by Forsythe, Nelson, Neumann and Wright (1992). Average daily prices on this market for the month of October 2008, and specifically the week in which our experiments were conducted, implied that the market probability of Obama winning was around 85%. Moreover, this was the prevailing sense of the market for at least 2 weeks prior to our experiments.[28] Indeed, the online Irish betting house, PaddyPower.com, was already paying out on over €1 million in pro-Obama bets as early as October 17! However, calibrating these inferred market probabilities to reflect the fact that most populations are risk averse on average would imply an even higher subjective probability.[29]

The prices from the Iowa Presidential Market on the popular vote share suggested that a 10% difference was possible. On the other hand, one striking feature of the two contracts, the "winner take all" contract and the "popular vote share" contract, is that the latter was poorly traded in comparison to the former as measured by volume. Of course, that could reflect a market that has found an equilibrium price, but it also could reflect a market in which there is too much uncertainty about the outcome for traders to feel safe making a bet. Additionally, correcting for the risk preferences of the average market participant would imply a much lower probability, consistent again with what we report here.

[26] The estimates for the subjective probability π refer to a non-linear transform in which we actually estimate the parameter κ and then convert κ to π using $\pi=1/(1+\exp(\kappa))$. Thus κ can vary between $\pm\infty$ and π is constrained to the open unit interval. To interpret these coefficients, $\kappa=0$ implies $\pi=½$, $\kappa>0$ implies $\pi<½$, and $\kappa<0$ implies $\pi>½$. The estimated subjective probabilities we report have been converted back from κ to π using this non-linear function and the "delta method" to correctly calculate standard errors (Oehlert (1992)). In addition, it is the linear function of κ that is constrained by this transform to be in the unit interval, not each element of that function. Thus, in Table B3 the constant term for π has a statistically significant coefficient of −0.7, which would violate that constraint if there were no covariates.

[27] For example, men and women might have different risk attitudes or subjective beliefs, and the mix of men and women could vary from treatment to treatment. This can occur even with randomization to treatment, particularly when considering a wide range of covariates.

[28] And the same was true generally, and consistently, for 9 months prior to the 2012 presidential election, even if the popular vote was relatively close.

[29] Fountain and Harrison (2011) examine many ways on which averages or medians from prediction markets might not reflect the average or median of the aggregate distribution of beliefs, apart from heterogeneity of risk attitudes.

5 Conclusions

We demonstrate how one can apply the theory of subjective probability elicitation by means of scoring rules. Our experimental design shows how one can pair different types of choice tasks to allow estimation of risk attitudes, which can then be used to condition the inferences from responses to the scoring rule tasks. Our method involves two simple types of choice tasks that allow us to identify not just the curvature of utility but also of probability weighting, and to control for these when inferring subjective beliefs. Our structural econometric model shows how maximum likelihood methods can then be used to estimate subjective probabilities while properly controlling for both the heterogeneous risk attitudes and for the statistical errors of the latter. We applied this approach to elicit subjective probabilities over four naturally occurring events from a sample of 140 subjects. We find that it is important to correct for risk attitudes both through the utility curvature and probability curvature. For these data we find that both utility functions and probability functions are concave, so that the corrections of probability reports work in opposite directions, at least for likely events. In addition, the magnitude of the utility correction is smallest close to 50%, while this is where the magnitude of the probability correction is the strongest. Once we correct for risk attitudes using our joint estimation approach we find no consistent significant difference between the QSR and LSR. This makes the LSR appealing due to the simplicity of explaining it to subjects.

Our results show that one has to be sensitive to the risk attitudes of subjects before drawing inferences about subjective probabilities from responses to scoring rules. One cannot just directly treat the response to the scoring rule as if it is a subjective probability, unless one is willing a priori to make striking assumptions about risk attitudes. Those assumptions are rejected in our data.

Quite apart from inferring the correct point estimate of subjective probability, uncertainty about risk attitudes affects the confidence that one should have in any point estimate. Even if subjects are "approximately risk neutral" on average, and the QSR is used, uncertainty about the precise level of risk attitudes should be properly reflected in uncertainty about inferences over subjective probabilities. Our analysis has demonstrated how to combine theory and econometrics to do just that. The choice task for eliciting subjective probabilities generates a point response that might appear to be quite precise by itself, but which is actually not a very precise estimate of the latent subjective probability when one properly accounts for uncertainty over risk attitudes.

Of course, although the estimation of subjective probabilities is an important objective in itself, the issue of how to best characterize subjective uncertainty, and attitudes towards it, involve deeper issues. Our estimation approach is clearly within the conventional Bayesian subjective probability framework of Savage (1971, 1972). That framework provides one obvious point of departure for criticisms of EUT based on ideas that subjective beliefs should not be represented by subjective probabilities. Exactly how one then models subjective beliefs is an open and important area of research (e.g., Smith (1969), Gilboa and Schmeidler (1989), Ghirardoto, Maccheroni and Marinacci (2004), Klibanoff, Marinacci and Mukerji (2005), Nau (2007) and Gilboa, Postlewaite and Schmeidler (2008). Further, it seems plausible that subjects exhibit some degree of "uncertainty aversion" in addition to traditional risk aversion when faced with making decisions about events that involve subjective probabilities rather than objective probabilities. We leave these issues to future research.

Acknowledgments We thank the U.S. National Science Foundation for research support under grants NSF/HSD 0527675 and NSF/SES 0616746. We are grateful to the Editor, a referee, John Duffy, J. Todd Swarthout, Peter Wakker, Randall Walsh and Nathaniel Wilcox for discussions and helpful comments. Complete instructions, additional estimation results, and details of the econometric specification are provided in online appendices.

References

Baron-Cohen, S. (2003). *The essential difference: The truth about the male and female brain*. New York: Basic Books.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science, 9*, 226–232.

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica, 74*, 1579–1601.

Costa-Gomes, M. A., & Weizsäcker, G. (2008). Stated beliefs and play in normal-form games. *Review of Economic Studies, 75*, 729–762.

Croson, R. (2000). Thinking like a game theorist: Factors affecting the frequency of equilibrium play. *Journal of Economic Behavior and Organization, 41*, 299–314.

de Finetti, B. (1937). La PRÉVISION: Ses Lois Logiques, Ses Sources Subjectives, Annales de l'Institut Henri Poincairé, 7, 1–68; English translation as: Foresight: Its logical laws, its subjective sources. In H. E. Kyburg & H. E. Smokler (Eds.), *Studies in subjective probability*. Robert E. Krieger: Huntington, NY.

de Finetti, B. (1970). Logical foundations and measurement of subjective probability. *Acta Psychologica, 34*, 129–145.

Dufwenberg, M., & Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior, 30*, 163–182.

Epstein, R. A. (1977). *The theory of gambling and statistical logic*. San Diego: Academic.

Evans, W. N., & Viscusi, W. K. (1993). Income effects and the value of health. *Journal of Human Resources, 28*, 497–518.

Forsythe, R., Nelson, F., Neumann, G. R., & Wright, J. (1992). Anatomy of an experimental political stock market. *American Economic Review, 82*, 1142–1161.

Fountain, J., & Harrison, G. W. (2011). What do prediction markets predict? *Applied Economics Letters, 18*, 267–272.

Ghirardoto, P., Maccheroni, F., & Marinacci, M. (2004). Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory, 118*, 133–173.

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

Gilboa, I., Postlewaite, A. P., & Schmeidler, D. (2008). Probability and uncertainty in economic modeling. *Journal of Economic Perspectives, 22*, 173–188.

Grether, D. M. (1992). Testing Bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization, 17*, 31–57.

Hanson, R. (1996). Correction to McKelvey and Page, 'public and private information: An experimental study of information pooling.'. *Econometrica, 64*, 1223–1224.

Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers, 107–119*.

Hao, L., & Houser, D. (2012). Belief elicitation in the presence of naïve respondents: An experimental study. *Journal of Risk and Uncertainty, 44*(2), 161–180.

Harrison, G. W., & Rutström, E. E. (2008). Risk Aversion in the Laboratory. In J. C. Cox & G. W. Harrison (Eds.), *Risk aversion in experiments, 12*. Bingley: Emerald Group Publishing Limited.

Harrison, G.W., Martínez-Correa, J., Swarthout, J.T. & Ulm, E. (2012). Scoring rules for subjective probability distributions, working paper 2012-10, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.

Harrison, G. W., Martínez-Correa, J., & Swarthout, J. T. (2013a). Inducing risk neutral preferences with binary lotteries: A reconsideration. *Journal of Economic Behavior & Organization, 94*(2013), 145–149.

Harrison, G. W., Martínez-Correa, J., & Swarthout, J. T. (2013b). Eliciting subjective probabilities with binary lotteries. *Journal of Economic Behavior & Organization, 101*(2014), 128–140.

Harstad, R. M. (2000). Dominant strategy adoption and bidders' experience with pricing rules. *Experimental Economics, 3*, 261–280.

Haruvy, E., Lahav, Y., & Noussair, C. (2007). Traders' expectations in asset markets: Experimental evidence. *American Economic Review, 97*, 1901–1920.

Hey, J. D., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica, 62*(1994), 1291–1326.
Holt, C. A. (1986). Scoring rule procedures for eliciting subjective probability and utility functions. In P. K. Goel & A. Zellner (Eds.), *Bayesian inference and decision theory: Essays in honor of Bruno de Finetti*. Amsterdam: North-Holland.
Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review, 92*, 1644–1655.
Holt, C. A., & Smith, A. M. (2009). An update on Bayesian updating. *Journal of Economic Behavior & Organization, 69*, 125–134.
Hurley, T. M., & Shogren, J. F. (2005). An experimental comparison of induced and elicited beliefs. *Journal of Risk and Uncertainty, 30*, 169–188.
Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2008). Scoring rules, generalized entropy, and utility maximization. *Operations Research, 56*, 1146–1157.
Kadane, J. B., & Winkler, R. L. (1987). De Finetti's method of elicitation. In R. Viertl (Ed.), *Probability and Bayesian statistics*. New York: Plenum.
Kadane, J. B., & Winkler, R. L. (1988). Separating probability elicitation from utilities. *Journal of the American Statistical Association, 83*, 357–363.
Karni, E. (1999). Elicitation of subjective probabilities when preferences are state-dependent. *International Economic Review, 40*, 479–486.
Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica, 77*, 603–606.
Karni, E., & Safra, Z. (1995). The impossibility of experimental elicitation of subjective probabilities. *Theory and Decision, 38*, 313–320.
Klibanoff, P., Marinacci, M., & Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica, 73*, 1849–1892.
Köszegi, B., & Rabin, M. (2008). Revealed mistakes and revealed preferences. In A. Caplin & A. Schotter (Eds.), *The foundations of positive and normative economics: A handbook*. New York: Oxford University Press.
Marschak, J. (1964). Actual versus consistent decision behavior. *Behavioral Science, 9*, 103–110.
Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science, 22*, 1087–1096.
McDaniel, T. M., & Rutström, E. E. (2001). Decision making costs and problem solving performance. *Experimental Economics, 4*, 145–161.
McKelvey, R. D., & Page, T. (1990). Public and private information: An experimental study of information pooling. *Econometrica, 58*, 1321–1339.
Nau, R. F. (2007). Extensions of the subjective expected utility model. In W. Edwards, R. Miles Jr., & D. von Winterfeldt (Eds.), *Advances in Decision Analysis: From Foundations to Applications*. New York: Cambridge University Press.
Nyarko, Y., & Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica, 70*(2002), 971–1005.
O'Hagen, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Hoboken, NJ: Wiley.
Oehlert, G. W. (1992). A note on the delta method. *The American Statistician, 46*, 27–29.
Offerman, T., Sonnemans, J., & Schram, A. (1996). Value orientations, expectations and voluntary contributions in public goods. *Economic Journal, 106*, 817–845.
Offerman, T., Sonnemans, J., van de Kuilen, G., & Wakker, P. P. (2009). A truth-serum for non-Bayesians: Correcting proper scoring rules for risk attitudes. *Review of Economic Studies, 76*, 1461–1489.
Plott, C. R., & Zeiler, K. (2005). The willingness to pay-willingness to accept gap, the 'endowment effect', subject misconceptions, and experimental procedures for eliciting valuations. *American Economic Review, 95*, 530–545.
Prelec, D. (1998). The probability weighting function. *Econometrica, 66*, 497–527.
Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization, 3*, 323–343.
Rutström, E. E. (1998). Home-grown values and the incentive compatible auction design. *International Journal of Game Theory, 27*, 427–441.
Rutström, E. E., & Wilcox, N. T. (2009). Stated beliefs versus empirical beliefs: A methodological inquiry and experimental test. *Games and Economic Behavior, 67*, 616–632.
Saha, A. (1993). Expo-power utility: A flexible form for absolute and relative risk aversion. *American Journal of Agricultural Economics, 75*, 905–913.
Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of American Statistical Association, 66*, 783–801.
Savage, L. J. (1972). *The foundations of statistics*. New York: Dover Publications.

Schotter, A., & Sopher, B. (2003). Social learning and coordination conventions in intergenerational games: An experimental study. *Journal of Political Economy, 111*, 498–529.

Shephard, G. G., & Kirkwood, C. W. (1994). Managing the judgemental probability elicitation process: A case study of analyst/manager interaction. *IEEE Transactions on Engineering Management, 41*, 414–425.

Smith, V. L. (1969). Measuring nonmonetary utilities in uncertain choices: The Ellsberg urn. *Quarterly Journal of Economics, 83*, 324–329.

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

Viscusi, W. K., & Evans, W. N. (1990). Utility functions that depend on health status: Estimates and economic implications. *American Economic Review, 80*, 353–374.

Viscusi, W. K., & Evans, W. N. (1998). Estimation of revealed probabilities and utility functions for product safety decisions. *Review of Economics and Statistics, 80*, 28–33.

Viscusi, W. K., & Evans, W. N. (2006). Behavioral probabilities. *Journal of Risk and Uncertainty, 32*, 5–15.

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