Scoring Rules for Subjective Probability Distributions

by

Glenn W. Harrison, Jimmy Martínez-Correa, J. Todd Swarthout and Eric R. Ulm †

Forthcoming, *Journal of Economic Behavior & Organization*

November 2016

ABSTRACT.

Subjective beliefs are elicited routinely in economics experiments. However, such elicitation often suffers from two possible disadvantages. First, beliefs are recovered in the form of a summary statistic, usually the mean, of the underlying latent distribution. Second, recovered beliefs are biased significantly due to risk aversion. We characterize an approach for eliciting the entire subjective belief distribution that is minimally biased due to risk aversion. We offer simulated examples to demonstrate the intuition of our approach. We also provide theory to formally characterize our framework. And we provide experimental evidence which corroborates our theoretical results. We conclude that for empirically plausible levels of risk aversion, one can reliably elicit most important features of the latent subjective belief distribution without undertaking calibration for risk attitudes providing one is willing to assume Subjective Expected Utility.

† Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA (Harrison); Department of Economics, Copenhagen Business School, Denmark (Martínez-Correa); Department of Economics, Andrew Young School of Policy Studies, Georgia State University, USA (Swarthout); and Department of Risk Management & Insurance, Robinson College of Business, Georgia State University, USA (Ulm). E-mail contacts: gharrison@gsu.edu, jima.eco@cbs.dk, swarthout@gsu.edu and eulm@gsu.edu. Harrison is also affiliated with the School of Economics, University of Cape Town. e are grateful to David Gonzalez Larrahondo for research assistance, and to the Society of Actuaries and the Center for Actuarial Excellence Research Fund for partial financial support.

1. Introduction

The experimental economics literature contains many instances of subjective belief elicitation. The elicitation of beliefs provides valuable insight into the latent decision-making process. These elicitation exercises have, almost always, involved estimating a summary statistic of the true underlying subjective belief distribution – usually the expected value. For some applications, a *summary statistic* of the latent belief distribution may be sufficient. However, there are other applications for which knowledge of the actual latent belief *distribution* itself would be beneficial. With this in mind, we characterize the properties of a procedure for eliciting entire subjective belief distributions. Through the use of simulation, theory and an experiment, we demonstrate several attractive properties of our procedure.[1]

It is well known that risk aversion can dramatically affect the incentives to correctly report the true subjective *probability* of a *binary* event.[2] To address this inferential problem, one can "calibrate" inferences about true subjective probabilities from elicited subjective probabilities, recognizing the incentives that risk averse agents have to report the same probability for the two outcomes and reduce the variability of payoffs from the scoring rule.[3] Or one must use relatively more complex scoring rules that "risk-neutralize" the agent.[4] Or one must eschew the use of any incentives for truthful elicitation.[5]

---

[1] The theoretical development of the specific scoring rule procedure we consider is due to Matheson and Winkler [1976]. Assuming risk neutrality, it has been used in the experimental literature in economics by Offerman, Sonnemans and Schram [1996], and in psychology by Moore and Healy [2008] and Merkle and Weber [2011], *inter alia*. Our general contribution is to propose a general characterization of the validity of this procedure under SEU, not assuming risk neutrality. We also provide an intuitive operationalization of the elicitation procedure. Harrison and Ulm [2016] extend our results to show how to recover latent beliefs when probability weighting behavior is identified for the individual.

[2] See Winkler and Murphy [1970], Savage [1971; p. 785] and Kadane and Winkler [1988].

[3] See Fiore, Harrison, Hughes and Rutström [2009], Offerman, Sonnemans, van de Kuilen and Wakker [2009], Andersen, Fountain, Harrison and Rutström [2014] or Antoniou, Harrison, Lau and Read [2015][2016].

[4] See Smith [1961], Grether [1992], Köszegi and Rabin [2008; p.199], Karni [2009], Holt and Smith [2009] and Harrison, Martínez-Correa and Swarthout [2014].

[5] Delavande, Gineé and McKenzie [2001; p. 156] make the case for not bothering about incentives. Referring to studies in developing countries that have all been hypothetical, they argue that "even without

Given the bias that risk aversion presents when eliciting a subjective *probability*, it would be reasonable to think that a similar bias exists when eliciting a subjective *distribution*. In fact, we characterize the implications of the general SEU case of a risk averse agent when facing the popular Quadratic Scoring Rule (QSR) over *continuous* events,[6] and find that the biasing effect of risk aversion is severely *reduced* as compared to the case of eliciting the probability of a binary event. On reflection, the intuition is simple. Under SEU, risk aversion causes a "left or right distortion" in reports compared to latent beliefs over binary events, but only causes a "flattening distortion" in reports compared to latent beliefs over continuous events, spread and diffused over two or more reports in finite implementations. These two very different *patterns* of distortion are the result of preferences for payoff smoothing over (binary or continuous) events.

For empirically plausible levels of risk aversion, which we quantify below, our theoretical results imply that one can reliably elicit the most important features of the latent subjective belief distribution without undertaking calibration for risk attitudes. In particular, as explained in the example, under relatively weak conditions we can recover the mean of the subjective distribution.[7] The mean is of particular interest to economists, since it corresponds to the single subjective probability to which subjective belief distributions are reduced under SEU by application of the Reduction of Compound Lotteries axiom. Therefore one could test, for example, if deviations from SEU theory, as modeled by ambiguity aversion theories, are due to violations of this axiom by comparing if choices under ambiguity are consistent with the elicited mean of the distribution.

---

payment, the answers received from such questions appear reasonable, and as such, there seems to have been a *de facto* decision that payments are not needed." We do not know what "reasonable" might possibly mean when it comes to subjective beliefs.

[6] We use the expression "continuous events" to refer to events with a continuous probability distribution over a continuum of possible outcomes, i.e. with continuous support. These continuous events are continuous intervals of the support of the distribution and therefore can be infinitesimally small, a property that will be used in some of our theoretical results.

[7] If one is only interested in eliciting the *mean* of a continuous subjective belief distribution, and willing to assume risk neutrality, then one can directly apply the QSR as shown by Costa-Gomes, Huck and Weizsäcker [2014]. In effect, this is the same as our procedure in which the subject is forced to allocate all "reports" to one interval, as explained below.

In section 2 we provide examples of eliciting subjective belief distributions and simulate varying levels of risk aversion in order to illustrate the minimal bias risk aversion has on reported belief distributions. Following from the intuition provided by the preceding examples, in section 3 we present theory to formally characterize our intuition and discuss the implications of the theory. In section 4 we report an experiment developed from some of the testable implications of the theory. And we conclude in section 5.

2. Examples

Let the decision maker report his subjective beliefs in a discrete version of a QSR for continuous distributions (Matheson and Winkler [1976]). We consider the QSR since it is the most popular scoring rule in use, but show that all theoretical results generalize to any proper scoring rule, which is any rule for which truthful reporting generates the highest expected payoff .

Partition the domain into K intervals, and denote as r_k the report of the likelihood that the event falls in interval $k = 1, ..., K$. Assume for the moment that the decision maker is risk neutral, and that the full report consists of a series of reports for each interval, $\{ r_1, r_2, ..., r_k, ..., r_K \}$ such that $r_k \geq 0 \; \forall k$ and $\sum_{i=1...K} (r_i) = 1$.

If k is the interval in which the actual value lies, then the payoff score is defined by Matheson and Winkler [1976; p.1088, equation (6)]:

$$S = (2 \times r_k) - \sum_{i=1...K} (r_i)^2$$

So the reward in the score is a doubling of the report allocated to the true interval, and the penalty depends on how these reports are distributed across the K intervals. The subject is rewarded for accuracy, but if that accuracy misses the true interval the punishment is severe. The punishment includes all possible reports, including the correct one.

Take some examples, assuming $K = 4$. What if the subject has very tight subjective beliefs and allocates all of the weight to the correct interval? Then the score is

$$S = (2 \times 1) - (1^2 + 0^2 + 0^2 + 0^2) = 2 - 1 = 1,$$

and this is positive. But if the subject has tight subjective beliefs that are wrong, the score is

$$S = (2 \times 0) - (1^2 + 0^2 + 0^2 + 0^2) = 0 - 1 = -1,$$

and the score is negative. So we see that this score would have to include some additional "endowment" to ensure that the earnings are positive; this is a point of practical behavioral significance, but is not important for the immediate theory. Assuming that the subject has very diffuse subjective beliefs and allocates 25% of the weight to each interval, the score is less than 1:

$$S = (2 \times \tfrac{1}{4}) - ((\tfrac{1}{4})^2 + (\tfrac{1}{4})^2 + (\tfrac{1}{4})^2 + (\tfrac{1}{4})^2) = \tfrac{1}{2} - \tfrac{1}{4} = \tfrac{1}{4} < 1.$$

So the tradeoff from the last case is that one can always ensure a score of ¼, but there is an incentive to provide less diffuse reports, and that incentive is the possibility of a score of 1.

To ensure complete generality, and avoid any decision maker facing losses, allow some endowment, α, and scaling of the score, β. We then get the following scoring rule from the report $\{ r_1, r_2, ..., r_k, ..., r_K \}$ when the true event is in interval k:

$$\alpha + \beta \, [\, (2 \times r_k) \; - \; \Sigma_{i=1...K} \, (r_i)^2 \,],$$

where we initially assumed $\alpha=0$ and $\beta=1$. We can assume $\alpha>0$ and $\beta>0$ to get the payoffs to any positive level and units we want. Let p_k represent the underlying, true, latent subjective probability of an individual for an outcome that falls into interval k.

Figures 1, 2 and 3 illustrate this scoring rule for the case in which $K = 10$, $\alpha = \beta = 25$, and we assume a subjective expected utility maximizer with a CRRA utility function $u(w) = w^{1-\rho}/(1-\rho)$ such that $\rho = 0$ denotes risk neutrality and $\rho > 0$ risk aversion.[8] Figure 1 shows the simplest case in

[8] Utility is defined solely over the income generated by the scoring rule. If utility is event-dependent then one must assume away any effects of the subjective outcome on initial wealth (Kadane and Winkler [1988], Karni and Safra [1995]). In our experiment this is natural, since subjects are betting on the outcome of a draw from an urn that has no connection to events outside the lab, other than the income these bets might generate. In field applications of these scoring rules this assumption might not be so natural. For instance, one might be eliciting beliefs about housing prices from somebody that already owns a house, so that the possible events affect the value of the initial endowment the individual has before any income from the scoring rule. Or preferences themselves might be state-dependent, quite apart from any effect on the arguments of the utility function: different health outcomes, over which one might naturally have subjective beliefs, might affect the utility associated with given endowments. Finally, scoring rules might be embedded in a competitive environment in which performance relative to others becomes a factor. This can lead to an additional

which the true subjective distribution is symmetric. The histogram shows the true distribution, and the black "droplines" show the optimal report. Under risk-neutrality, the top left panel of Figure 1 shows that the individual truthfully reports the true subjective distribution. Looking at the additional panels in Figure 1, we see the trend toward the reported distribution being a flattened version of the subjective distribution as ρ increases from 0 up to 3. Also apparent is the complete absence of any reports for outcomes 1, 7, 8, 9 and 10, which have no subjective density.

Maintaining the true subjective belief distribution from Figure 1, we now consider the more realistic case in which the agent is risk averse at a parameter value typical in the laboratory (Harrison and Rutström [2008]). An individual with relative risk aversion of $\rho = 0.65$, for instance, under-reports the true subjective probability for outcome 4 ($r_4 = 0.356 < p_4 = 0.4$). Such an individual over-reports the true subjective probability for outcomes 3 and 5 ($r_3 = r_5 = 0.207 > p_3 = p_5 = 0.2$), but the distortion is barely noticeable. The over-reporting for outcomes 2 and 6, however, is noticeable ($r_2 = r_6 = 0.115 > p_2 = p_6 = 0.1$). Since the extent of the reporting deviations are the same on either side of the mode, and the true distribution is symmetric, the average of the reported distribution would always equal the average of the true distribution when rounded to the interval widths used in a specific application.

Figure 2 considers the case of an asymmetric, unimodal subjective distribution, and varying levels of risk aversion. For relative risk aversion level $\rho > 0$, the true probabilities for outcomes 6 and 5 are under-reported, and for outcomes 4 and 3 are over-reported. Again, there are no reports for outcomes that have no subjective density.

Finally, using the parameters and beliefs from Figure 2, Figure 3 shows how the average of the reported distribution deviates from the average of the true subjective distribution in the unimodal, asymmetric case. For a wide range of risk attitudes observed in the same experimental

distortion of reports (Lichtendahl and Winkler [2007]).

context that we would undertake these belief elicitations ($\rho < 1$), we find the difference to be less than a percentage point. Of course, there is no point showing comparable figures for the symmetric distributions, since in that case there is no difference at all when rounded to the interval widths used in a specific application.

The preceding discussion used numerical simulations to provide visual and descriptive evidence of our results. We now formalize our results with the theory in the following section.

3. Theory

We focus on the discrete case, in part for expository reasons, but also because this is the interesting case in terms of operational scoring rules. All proofs for the discrete case are in Appendix A. The proofs for the continuous case are similar, and collected in Appendix B in the Online Supplement. Again, unless otherwise noted, the following proofs assume that the decision maker is a SEU maximizer. Proposition 1 assumes symmetric subjective distributions; Propositions 2, 3, 4, 5 and 6 and the Lemmata do *not* assume symmetric subjective distributions, nor do they assume that the distribution is even unimodal.[9]

Lemma 1: Let p_k represent the underlying subjective probability of an individual for outcome k and let r_k represent the reported probability for outcome k in a given scoring rule. Let $w(k) = \alpha + \beta 2r_k - \beta \sum_{i=1 \dots K} (r_i)^2$ be the scoring rule that determines the wealth if state k occurs. If the individual has a utility function $u(w)$ that is continuous, twice differentiable, increasing and concave and maximizes expected utility over actual subjective probabilities, the actual and reported probabilities must obey the following system of equations:

[9] On the other hand, for many applications we have observed symmetry to be a plausible empirical assumption, providing one does not select bin intervals that lead to reports bunched at extremes (e.g., imagine eliciting beliefs about the future paths of nominal sovereign interest rates, when they are historically close to zero). Appendix E in the Online Supplement documents this empirical claim for our data, and suggests a general methodology for checking in specific applications. The plausibility of symmetry is likely to vary with the belief being elicited, but it is valuable to have a simple check for the plausibility of assuming it if one wants to apply Proposition 1.

$$p_k \times \partial u/\partial w \,|_{w=w(k)} - r_k \times E_p[\partial u/\partial w] = 0, \forall\, k = 1,\ldots, K \quad (1)$$

Lemma 2: Under the assumptions of Lemma 1, let $\varepsilon_k = r_k - p_k$ be the deviation between the reported and actual subjective probabilities for outcome k. Then

$$\varepsilon_k = p_k \times \{\partial u/\partial w \,|_{w=w(k)} - E_p[\,\partial u/\partial w\,]\}/E_p[\,\partial u/\partial w\,], \forall\, k=1,\ldots, K. \quad (7)$$

Lemma 3: Assume an individual who has a continuous, twice differentiable utility function $u(w)$ that is increasing in random wealth and who is also risk averse (i.e., $\partial^2 u/\partial^2 w < 0, \forall\, w$). If $p_i = p_j$ for some i and j, then $r_i = r_j$. This result does not hold for risk loving individuals.

Proposition 1: For the risk-averse individual in Lemma 3, if the underlying subjective distribution is symmetric then the mean of the reported distribution is equal to the mean of the actual subjective distribution.

Lemma 4: The converse of Lemma 3. Assume an individual with a continuous, differentiable utility function $u(w)$, where risk aversion is not necessary in this case. If $r_i = r_j$ for this individual, then $p_i = p_j$.

Proposition 2: For the individual in Lemma 4, if the reported distribution is symmetric then the mean of the reported distribution is equal to the mean of the actual subjective distribution.

Proposition 3: Assume an individual with a continuous, twice differentiable utility function $u(w)$ that is increasing in w. The individual reports probability $r_k = 0$ if and only if the true subjective probability of the individual for state k is $p_k = 0$.[10]

Proposition 4: A risk-averse individual has a reported probability distribution that approaches a uniform distribution over those states where $p_k > 0$ in the following sense: There exists a value p^* for this individual such that if $p_k > p^*$ then $p_k > r_k > p^*$ and if $p_k < p^*$ then $p_k < r_k < p^*$. A risk-loving agent reverses all the conditions.

---

[10] Proposition 3 also applies to models of decision-making under risk that allow for continuous, weakly-monotonic probability weighting functions with fixed points at 0 and 1.

Proposition 5: An individual with sufficiently high risk aversion will have a reported probability arbitrarily close to p^*.

Proposition 6: The following relationship exists between means of the reported and actual subjective distributions up to the first order approximation: If $u(w) = w + \delta \times u^*(w)$, then for any random variable y, $E_r[y] - E_p[y] \approx \delta \times Cov_p [\partial u/\partial w, y]$.

Proposition 7: The following relationship exists between means of the reported and actual subjective distributions up to the first order approximation: If $u(w) = w + \delta \times u^*(w)$, then for any random variable y, $E_r[y] - E_p[y] \approx Cov_r [\partial u/\partial w, y]$.

Proposition 8: Lemma 1 generalizes to include all proper scoring rules. Hence all of the results that flow from Lemma 1 also generalize.

We summarize and restate the implications of this theory, some of which we then test:

1. The individual *reports* a positive probability for an event only if the individual has a positive subjective probability for the event. So if the individual believes that inflation will never fall below 1.5% per annum, we would never see the individual reporting that it would.
2. If an individual has the same subjective probability for two events, then the reported probabilities for the two events will also be the same if the individual is risk averse or risk neutral. So if the individual attaches a probability of 0.25 to the chance that inflation will be between 1% and 2%, and a probability of 0.25 to the chance that inflation will be between 4% and 5%, the reported probabilities for these two intervals will be the same as well (although typically not 0.25).
3. The converse is true for risk averse subjects, as well as for risk lovers. That is, if we observe two events receiving the same reported probability, we know that the true subjective probabilities are also equal, although not necessarily the same as the reported probabilities.

4. If the individual has a *symmetric* subjective distribution, then the reported mean will be the "same" as the true subjective mean, whether or not the subjective distribution is unimodal.[11] This result is exact when the interval widths are arbitrarily small, and is valid in terms of the rounded means when the interval widths are finite. Further, this result follows from the previous two results, and is of great significance for tests of theories about ambiguous events that stress the role of the Reduction of Compound Lotteries axiom over subjective belief distributions (e.g., Nau [2006] and Ergin and Gul [2009]). A testable implication of that axiom is that the individual behaves as if holding a subjective probability equal to the average of some subjective belief distribution. Hence if we simply assume symmetry of the true distribution, a relatively weak assumption in some settings, we can elicit that mean directly.
5. If the individual reports a symmetric distribution then their true subjective distribution is also symmetric. This result also follows from earlier results, but provides a practically useful way to check if the assumption of symmetric beliefs is plausible in particular instances.
6. The more risk averse an agent is, the more their reported distribution will resemble a uniform distribution defined on the support of their true distribution. In effect, risk aversion causes the individual to report a "flattened" version of their true distribution. Merkle and Weber [2011; p. 268] state this result intuitively, without proof.
7. It is possible to bound the effect of increased risk aversion on the difference between the reported distribution and true distribution. This result provides a characterization of the empirical finding that the reported distribution is "very close" to the true subjective distribution for a wide range of empirically plausible risk attitudes.

---

[11] We define a symmetric subjective distribution in the usual manner as being a continuous probability density function $p(s)$ for which there exists some ξ such that $p(\xi-s) = p(\xi+s)$ for all real s. It refers to the underlying, true subjective probability density function, not any "discretized" probability mass function over which we elicit reports.

8. All of these results for the QSR generalize to any proper scoring rule.

4. Experiment

The above theoretical results are meant to help apply and interpret empirical efforts to elicit subjective belief distributions. Many of the properties of the scoring rule cannot be directly tested, given that they refer to unknown subjective beliefs: *de opinio non est disputandum*. For instance, Proposition 3 is a valuable property, but to test it we would need to know that some individual attached zero subjective weight to some specific interval of events.

However, it is possible with a controlled laboratory experiment to offer some evidence in support of the claim that, for a risk-averse individual with symmetric subjective beliefs, the mean of the reported distribution is equal to the mean of the actual subjective distribution (Proposition 1). In the laboratory we can present a stimulus for beliefs which provides a basis for symmetric beliefs, and for which we know, by design, the true objective probability distribution implied by the stimulus. It is then a simple matter to compare that true stimulus with the average elicited belief, under the maintained assumption that there is no basis for subjective beliefs to be biased in comparison to the true stimulus. This assumption cannot be easily made outside of this laboratory environment. Of course, our conclusions are limited to these controlled stimuli.

Our experiment elicits beliefs from subjects over the composition of a bingo cage containing both red and white ping-pong balls. Subjects did not know with certainty the proportion of red and white balls, but they did receive a noisy signal from which to form beliefs. The subjects were told that there were no other salient, rewarded choices for them to make before or after they made their choices, avoiding possible confounds by having to assume the "isolation effect" if one were making many choices.[12]

---

[12] The "random lottery" payment protocol in which one asks the subject to make K>1 choices, and pick 1 of the K at random for payment at the end, requires that the Mixture Independence axiom applies, or

Our experimental design consisted of 8 laboratory sessions. A total of 184 participants were recruited from a general subject pool of undergraduates at Georgia State University. We implement two between-subjects treatments within sessions 1-4 so that both groups are presented with the same randomly chosen and session-specific stimulus, thus we are able to compare treatment effects while conditioning on a specific realized stimulus. In **treatment 10bin** we elicit subjective belief *distributions* about the true fraction of red balls in the bingo cage by using a generalized QSR with monetary outcomes. In this treatment the belief elicitation tool divides the possible proportions of red balls into 10 disjoint sets that we label bins,[13] and subjects can make bets on each bin according to their beliefs. In **treatment 2bin** we elicit subjective *probabilities* that a single red ball would be drawn from bingo cage by using the QSR with monetary outcomes. In this treatment the ball drawn from the Bingo cage can be red or not, and thus subjects can make bets only on those two events that are represented by two bins. This probability elicitation task is known to be one in which risk averse subjects would rationally and significantly distort their reports towards ½. By comparing elicited reports across treatments in these sessions, we can assess the practical significance of our claims about the weak effects of risk aversion on optimal reports under treatment **10bin**.[14] In sessions 5-8 we only conducted the **10bin** treatment, in exactly the same manner as the **10bin**

---

at least a violation of one of the two axioms that constitute it (i.e., the Reduction of Compound Lotteries axiom and the Compound Independence axiom). But then one cannot use those data to estimate models of decision-making behavior that assumes the invalidity of that axiom. The only reliable payment protocol in this case is to ask subjects to only make one choice, and pay them for it. See Harrison and Swarthout [2014] for discussion, including the literature evaluating the behavioral validity of the isolation effect and the theoretical explanation of the behavioral axioms needed for the lottery payment mechanism to work.

[13] We use the term "bin" as synonymous with the term "event" from standard probability theory. Specifically, we define a *bin* as a set of outcomes of a random process to which a probability measure is assigned. We elected to refer to *bins* in our experiment, thinking that this term would be more natural for our subjects.

[14] Our **2bin** treatment is *formally* the same as eliciting the subjective fraction of the true distribution, rather than a single-draw realization. It could also have been operationalized by giving subjects the 10bin interface and restricting them to allocate all 100 tokens to only one interval. Whether these different ways of operationalizing the task are behaviorally identical is another matter. We see no reason to expect them to be different, but recognize that funny things can always happen in the lab.

treatment in sessions 1-4. Table 1 presents the number of subjects by treatment in each of the 8 sessions.

Each session was conducted in the manner described below. Upon arrival at the laboratory, each subject drew a number from a box which determined random seating position within the laboratory. After being seated and signing the informed consent document, subjects were given printed introductory instructions and allowed sufficient time to read these instructions (Appendix C in the Online Supplement provides complete subject instructions). Then a Verifier was selected at random among the subjects solely for the purpose of verifying that the procedures of the experiment were carried out according to the instructions. The Verifier was paid a fixed amount for this task and did not participate in the decision-making task.

Each subject was assigned to one of the two groups depending on whether their seat number was even or odd. One of the treatment groups was then taken out of the lab for a few minutes, always under the supervision of an experimenter. The other group remained in the laboratory and went over the treatment-specific instructions with an experimenter. Simultaneously, subjects waiting outside were given instructions to read individually. Then the groups swapped places and the experimenter read the treatment-specific instructions designed for the other group. Once all instructions were finished, and both groups were brought together in the room again, and we proceeded with the remainder of the experiment.

We used two bingo cages: Bingo Cage 1 and Bingo Cage 2. Bingo Cage 1 was loaded with balls numbered 1 to 99 in front of everyone.[15] A numbered ball was drawn from Bingo Cage 1, but the draw took place behind a divider. The outcome of this draw was not verified in front of subjects

[15] When shown in public, Bingo Cages 1 and 2 were always displayed in front of the laboratory where everyone could see them. We also used a high resolution video camera to display the bingo cages on three flat screen TVs distributed throughout the laboratory, and on the projection screen at the front of the room. Our intention was for everyone to have a generally equivalent view of the bingo cages.

until the very end of the experiment, after their decisions had been made. The number on the chosen ball from Bingo Cage 1 was used to construct Bingo Cage 2 behind the divider. The total number of balls in Bingo Cage 2 was always 100: the number of red balls matched the number on the ball drawn from Bingo Cage 1, and the number of white balls was 100 minus the number of red balls. Since the actual composition of Bingo Cage 2 was only revealed and verified in front of everybody at the end of the experiment, the Verifier's role was to confirm that the experimenter constructed Bingo Cage 2 according to the randomly chosen numbered ball. Once Bingo Cage 2 was constructed, the experimenter put the chosen numbered ball in an envelope and affixed it to the front wall of the laboratory.

Bingo Cage 2 was then covered with a black blanket and placed on a platform in the front of the room. After subjects were alerted to pay attention, Bingo Cage 2 was then uncovered for subjects to see, spun for 10 turns, and covered again. This visual display was the information that each subject received. Subjects then made their decisions based on this information about the number of red and white balls in Bingo Cage 2. After decisions were made, subjects completed a non-salient demographic survey. Immediately after, earnings were determined. To resolve payments in treatment 2bin, the experimenter drew a ball from Bingo Cage 2. The sealed envelope was then opened and the chosen numbered ball was shown to everyone, and the experimenter publicly counted the number of red and white balls in Bingo Cage 2.

A computer interface was used to present to subjects the belief elicitation tasks and to record their choices, allowing them to allocate tokens to reflect their subjective beliefs. Figure 4 presents the interface used for the distribution elicitation treatment (treatment **10bin**). The interface implements the QSR discussed earlier, with $\alpha=\beta=25$. Subjects could move the sliders at the bottom of the screen interface to re-allocate the 100 tokens as they wished, ending up with some distribution as shown in Figure 5. The instructions explained that they could earn up to $50, but only by allocating all 100 tokens to one interval *and* that interval containing the true percent: if the true percent was just outside the selected interval, they would in that case receive $0.

The stimulus, the number of red balls in Bingo cage 2, was different in each session since we wanted the true number of red balls to be generated in a credible manner, to avoid subjects second-guessing the procedure. This credibility comes at the risk that the stimulus is extreme and uninformative: if there had been only 1 red ball, or 99 red balls, we would not have generated informative data. As it happens, we had a good variety of realizations over the 4 sessions.

A. Eliciting Belief Distributions

Consider the subjective belief distributions elicited with the generalized QSR in treatment 10bin. We provide evidence consistent with our claims that, given the risk aversion observed in our subject population, the generalized QSR can directly elicit important features of subjective probability distributions. We provide some illustrative pictures and then formal statistical tests of the main hypotheses.

As a preliminary, we note that we have independent evidence that subjects from our population do "robustly" exhibit risk aversion over stakes comparable to those used in the present experiment: Holt and Laury [2002][2005], Harrison and Swarthout [2014], Harrison, Martínez-Correa and Swarthout [2015] and Harrison and Ng [2016]. All of these other experiments were run in the same laboratory with the same undergraduate population as our present experiment. Thus any correspondence with the predictions of Proposition 1 is not due to the risk neutrality of the subjects over these stakes.

To illustrate the data, Figure 6 presents the elicited beliefs pooled over the 15 subjects in the first session, in which the true percent was 69%. Of course there is some dispersion in beliefs, since the stimulus was deliberately designed not to provide exact information (unless, by unfortunate chance, the number of red balls was extreme). As it happens, the average of this elicited distribution is 71% and very close to the true proportion of red balls.[16]

Figure 7 reports the results across all sessions. With one exception, the elicited averages closely track the true averages. Again, the maintained joint hypothesis that allows us to view this as evidence for the truthful elicitation of subjective belief distributions is that subjects behave consistently with SEU *and* that their subjective belief distributions are distributed around the true population average that provides the common stimulus they all observe.

The clear exception in Figure 7 is session 7, in which the true number of red balls was 11% and the elicited average was 25%. This disparity is due to three outliers; we believe *a priori* these subjects did not understand the task. One subject allocated 36 tokens to the interval for 81% to 90%, and 64 tokens to the interval for 91% to 100%. It is possible this subject was confused as to whether he was betting on red or white. If this subject is removed, the average becomes 19%. Then there were two subjects who exhibited some degree of confusion, although less extreme than the first outlier.[17] If these two are also removed, the average becomes 16%, close to the true number of red balls. Of course one is always wary claiming that a subject is an outlier, although every behavioral economist knows that such subjects exist, and occasionally even in clusters like this.

We can formally statistically test the hypothesis that the elicited averages in Figure 7 are equal to the true percent by estimating an interval regression model in which the intervals are the bin

[16] The average is estimated using an interval regression model with no covariates. Hence the dependent variable is literally the interval selected by the subject, and the weight on that interval is the number of tokens allocated to the interval. Since the latent variable is bounded between 0% and 100%, we use an interval regression model assuming a Beta distribution rather than a Normal distribution.

[17] One of these subjects allocated roughly 10 tokens to each and every interval, which could have been due to not properly seeing the visual stimulus (e.g., inattention or some form of vision problem). The other subject allocated roughly 10 tokens to each interval below 50%, 28 tokens to the interval for 71% to 80%, and small numbers of tokens for other intervals greater than 50%.

"labels" in Figures 4 and 5, and the tokens allocated to each bin are frequency weights for each subject. We also cluster the standard errors on each subject. If we estimate this model with only a constant term and no covariates, we can directly test the hypothesis that the estimate of the constant term is equal to the true percent. For session 1 we estimate the mean to be 0.71 with a 95% confidence interval between 0.66 and 0.77, and a *p*-value of 0.42 on the null hypothesis that it is equal to 0.69. Hence we cannot reject the null hypothesis that average elicited beliefs are statistically significantly equal to the true percent. Table 2 shows comparable estimates and *p*-values for each session.

Table 2 has only one surprise, compared to the conclusions we would draw from Figure 7. This is in session 3, in which the true value was 0.10 and the estimated mean is 0.12, but the *p*-value is only 0.002. The reason for this very simple: we have a *very* precise estimate of the mean, with a standard error of only 0.006 and a 95% confidence interval between 0.11 and 0.13. Here we have a familiar conflict between statistical significance and economic significance. The difference between 0.10 and 0.12 is small in economic terms, but happens to be statistically significant. Moreover, the true value of 0.10 essentially straddles bin #1 and bin #2 in our elicitation tool, causing many subjects to allocate roughly 50% of their tokens to each of these bins. The bottom panel of Table 2 shows the effect of allowing for individual heterogeneity in the structural parameter that affects the variance of elicited distributions (conditional on the mean).[18] The effect of allowing for heterogeneity in variance is to slightly widen the 95% confidence interval on the estimate of the mean, resulting in a *p*-value on the hypothesis test of 0.09.

Session 7 also has a low *p*-value of 0.03 in Table 2 on the hypothesis test, but this is no surprise given the outliers noted above. If they are removed the *p*-value becomes 0.10. If we leave

[18] Formally, this is the "precision" parameter φ. The variance of the inferred latent belief distribution is equal to $\mu(1+\mu)/(1+\varphi)$, where μ is the mean.

them in, and allow for individual heterogeneity in the variance, the *p*-value becomes 0.13 as shown in the bottom panel of Table 2.

Apart from these two cases, the evidence is clearly consistent with the claim that we have elicited belief distributions whose averages are close to the true means of the stimuli.

B. Eliciting Distributions Versus Probabilities

We conduct a test of the effect of risk aversion by comparing elicited beliefs *for the same physical stimulus* but using different scoring rules. As is well known, the QSR for *binary* events will elicit biased responses if the subject is risk averse: intuitively, the subject is drawn to report 50% so as to equalize earnings under each possible outcome, providing subjective beliefs are not degenerate.[19] In sessions 1 through 4, we compare the elicited belief distributions discussed above with elicited probabilities based on the QSR for binary events, as shown in Figure 8.[20] In this case the binary event was a single draw from Bingo Cage 2 containing the red and white balls. Although all subjects within a given session are presented with the same physical stimulus, the two groups of

[19] A potential confound is the initialization of the token allocation at 50 tokens to each two bins in Figure 8. This might have generated an "anchoring effect" in which subjects were drawn to report 50 tokens to each bin for reasons other than risk aversion. Online appendix D explains why we deliberately chose this initialization and, more importantly, reviews evidence from Andersen, Fountain, Harrison and Rutström [2014] that there is no systematic or statistically significant evidence of anchoring on initial allocations in this type of task.

[20] A referee makes the astute observation that there are two, potentially confounding, issues to be considered. First, if one is not careful in selecting the K discrete intervals for the distribution elicitation task one could find no difference between the elicited distribution and the elicited probability since the former would not have much discriminatory power. For instance, if K=2 one would not be able to tease the two apart. In the experiments reported here it is an easy matter to select 10 intervals that are *a priori* informative, but in general applications this might require some care. A practical, field example is provided by Harrison and Phillips [2014], where the intervals for subjective belief elicitation of 11 global financial risks are selected by first estimating a simple econometric time series model and using it to forecast a 95% confidence interval from the data. Selecting intervals for the belief elicitation that span that interval provide informative comparisons of the forecasts of a statistical model and the subjective beliefs of Chief Risk Officers, which is the objective of the exercise from a risk management perspective. Second, even if one is careful to select, say, 10 intervals to span a wide range of responses, there is no guarantee that subjects will report more than 2 of those intervals. Appendix F in the Online Supplement demonstrates that this potential concern is not applicable in our case, since we observe over 70% of reports in the **10bin** treatments spanning more than 2 intervals.

subjects face different tasks: in the binary case the individual is betting over their subjective perception of an order statistic, and in the 10-event case the individual is betting over their subjective perception of a sufficient statistic of the population. Nonetheless, for the sessions in which the stimuli were not close to 50% already, sessions 1 through 3, we observe in Figure 9 a striking tendency for the reports using the binary scoring rule to be closer to 50% than to the true proportion. This is perfectly consistent with our predictions, since risk aversion has a significantly biasing effect for the binary scoring rule, and virtually none for the 10-event scoring rule.[21]

We now undertake a formal statistical test of the hypothesis that the average of the **2bin** and **10bin** treatments are the same. In this case we add a binary covariate to the interval regression model to identify the **2bin** responses, and test the statistical significance of that variable's coefficient on the mean of the latent subjective belief distribution.[22] For sessions 1, 2 and 3 the *p*-values on these tests are each less than 0.0001, and for session 4 the *p*-value is 0.53. These results are consistent with the conclusions drawn from Figure 9.

5. Conclusions

These results provide strong support for the use of practical methods for eliciting subjective belief *distributions* over *continuous* events. Contrary to the case in which one elicits subjective *probabilities* over *binary* events, there is *a priori* and empirical support for not needing to adjust or de-

[21] An astute referee notes that while this striking difference in the 2bin distortion and 10bin distortion is in the *direction* predicted by risk aversion, the *size* of the distortion in the 2bin treatment is larger than can be explained by risk aversion due to diminishing marginal utility alone. In fact, we conjecture that it is due to the sharp effects of probability weighting, which also can generate risk aversion. The quantitative importance of probability weighting in the 2bin case is discussed and demonstrated by Andersen, Fountain, Harrison and Rutström [2014] in closely comparable experiments and samples. Harrison and Ulm [2016] demonstrate the same quantitative importance in the 10bin elicitation procedure considered here, by demonstrating how to recover latent beliefs when individuals probability weight, and showing that the distortions can be significant (e.g., their Figures 4, 10 and 12).

[22] The responses from the **2bin** treatment provide non-interval, point data, but these can be pooled with the interval data from the **10bin** treatment.

bias the reports for continuous events on account of risk aversion.[23]

Our findings are limited to agents who are assumed to follow SEU. One would ideally also like to have comparable procedures for eliciting whole distributions under alternative theories of decision making under risk, uncertainty or ambiguity, but this is a challenging task. Most popular alternatives to SEU allow for probability weighting behavior to occur, leading to decision weights that are non-additive (or directly assume non-additive decision weights). This adds a fundamental identification problem when trying to infer subjective beliefs. One can solve it, to some degree, by using methods developed for inferring subjective probabilities for binary events[24], and undertaking multiple elicitations and inferences to slice up the subjective distribution. For instance, instead of eliciting the complete distribution for the percentage return on the Standard and Poors Index in one year, one could elicit the probability that it is below -5%, between -5% and 0%, between 0% and 5%, and then that it is greater than 5%. This information could be then used directly as a coarse subjective belief distribution for the individual, or as the basis for some inferred parametric distribution.[25] However, it obviously adds several "chained" elicitation tasks, and requires calibration of each inference to risk attitudes or probability weighting behavior, depending on what model of behavior is assumed. We see our results as complementary to this approach: at the cost of assuming

[23] This is not the same as eliciting a series of binary subjective probabilities and "knitting together" an elicited subjective belief distribution. Our approach is to elicit the distribution in one task, not in a number of independent tasks. Undertaking a series of binary elicitations runs the risk of order effects, or the risk of elicited probabilities not summing to 1. It is also much harder to correctly estimate standard errors for the inferred latent distribution when making a series of independent inferences about binary slices of the underlying distribution. Of course, in future work it might be interesting to compare the consistency of elicited distribution from one task with constructed distribution from a series of binary elicitations.

[24] For example, Offerman, Sonnemans, van de Kuilen and Wakker [2009], Karni [2009] or Andersen, Fountain, Harrison and Rutström [2014], who each explicitly consider the case of probabilistically sophisticated non-SEU agents.

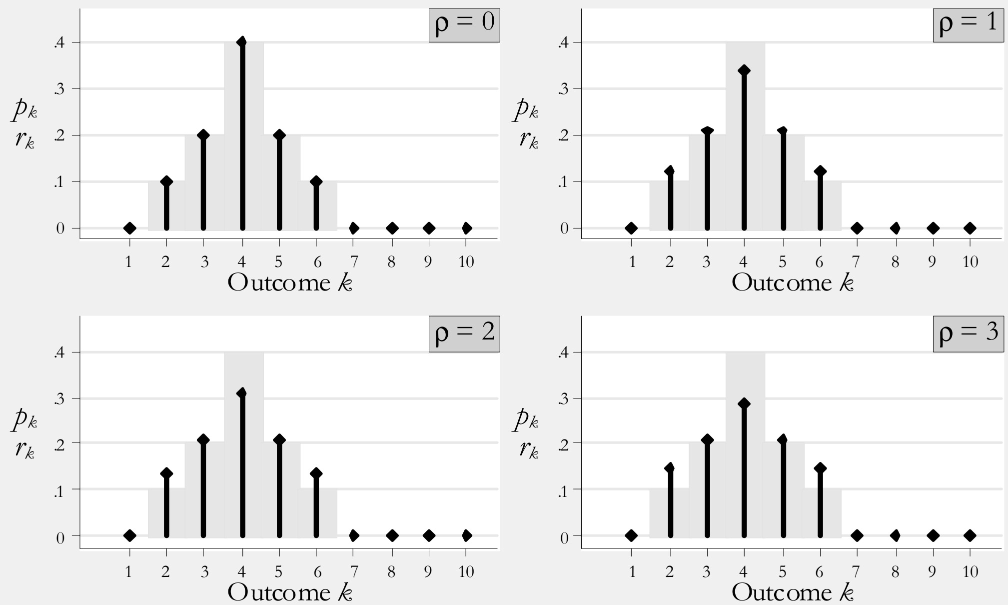
[25] An excellent example is the evaluation of the *Survey of Economic Expectations* responses on equity returns in Dominitz and Manski [2011; §2.2], although they do not consider the effects of risk aversion and probability weighting on inferences.

SEU, we argue that one can reliably infer whole subjective distributions in one task up to the discretization of the support space used in the elicitation task.[26] One task for future research is to compare elicited distributions with our approach and elicited "chained" distributions, where one can correct the latter for possible non-SEU behavior. Another task for future research is to use choices over objective risk to characterize subjects that are more likely to be EUT-consistent, and to make the weaker assumption that EUT-consistency implies SEU-consistency.[27]

[26] What is "one" task and what is "several tasks on one computer screen" is a semantic matter we do not want to be overly dogmatic about.

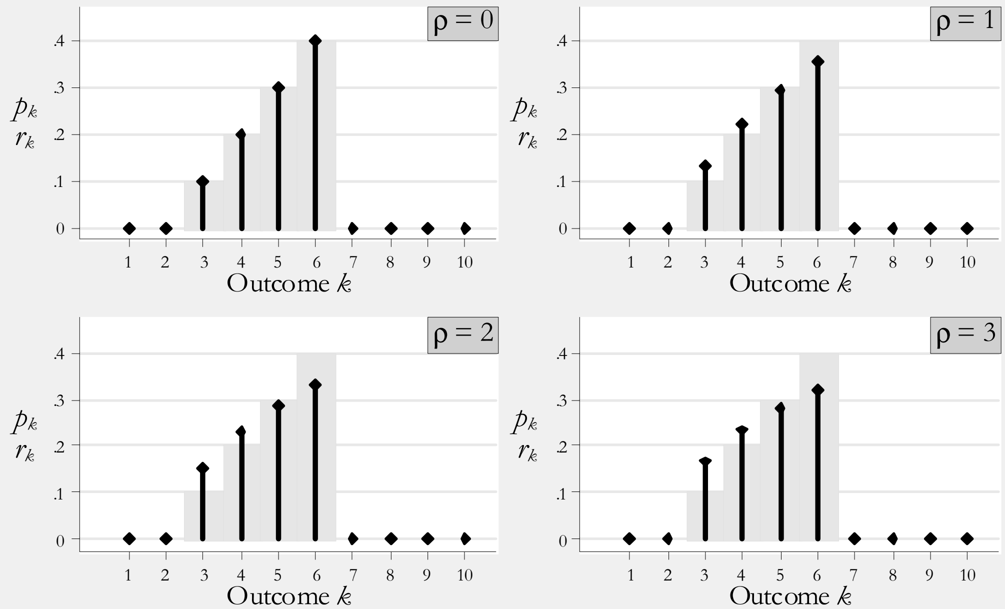
[27] To implement this robustness check one would add another series of tasks in which subjects make choices over objective risk so that one could test whether the subject was EUT-consistent, compared to some alternative such as Rank Dependent Utility (RDU), at a given significance level. It is then possible to see if the sub-sample of EUT-consistent subjects report different beliefs than the entire sample. Harrison [2014] implements this approach, and finds no statistical difference in elicited beliefs of RDU-consistent subjects. Of course, the maintained assumption here is that EUT-consistency is a useful measure of whether an individual is SEU-consistent.

Figure 1: Optimal Reports Assuming Unimodal, Symmetric Beliefs and Subjective Expected Utility



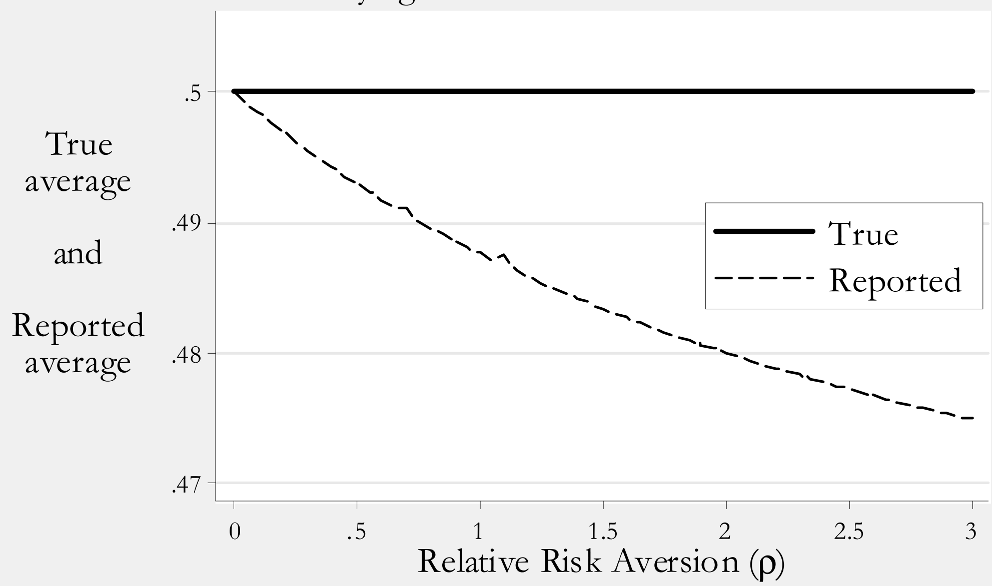
Note: This figure provides an example showing how a subjective probability distribution can deviate from the reported distribution under a scoring rule for the hypothetical case of a distribution with 10 possible outcomes. It compares a unimodal symmetric subjective probability distribution (gray bars) and implied optimal report (black "droplines") under the quadratic scoring rule $\alpha + \beta\,[\,(2 \times r_k) - \Sigma_{i=1...K}\,(r_i)^2\,]$, for the case in which $K = 10$, $\alpha = \beta = 25$. We further assume a subjective expected utility maximizer with a CRRA utility function $u(w) = w^{1-\rho}/(1-\rho)$ such that $\rho = 0$ denotes risk neutrality and $\rho > 0$ denotes risk aversion.

Figure 2: Optimal Reports Assuming Unimodal, Asymmetric Beliefs and Subjective Expected Utility



Note: This figure provides an example showing how a subjective probability distribution can deviate from the reported distribution under a scoring rule for the hypothetical case of a distribution with ten possible outcomes. It compares a unimodal asymmetric subjective probability distribution (gray bars) and implied optimal report (black "droplines") under the quadratic scoring rule $\alpha + \beta\,[\,(2 \times r_k) - \sum_{i=1...K} (r_i)^2\,]$, for the case in which $K = 10$, $\alpha = \beta = 25$. We further assume a subjective expected utility maximizer with a CRRA utility function $u(w) = w^{1-\rho}/(1-\rho)$ such that $\rho = 0$ denotes risk neutrality and $\rho > 0$ denotes risk aversion.

Figure 3: Difference Between True Average and Reported Average with Asymmetric Beliefs and Varying Relative Risk Aversion



Note: This figure shows how the average of the reported distribution deviates from the average of the true subjective distribution in the unimodal, asymmetric case. We use the same parameters used in Figure 2 with the scoring rule $\alpha + \beta\ [\ (2 \times r_k)\ -\ \sum_{i=1...K} (r_i)^2\ ]$, for the case in which $K = 10$, $\alpha = \beta = 25$ and a subjective expected utility maximizer with a CRRA utility function $u(w) = w^{1-\rho}/(1-\rho)$ such that $\rho = 0$ denotes risk neutrality and $\rho > 0$ denotes risk aversion. The bold line shows the assumed mean of the subjective distribution and the dotted line shows how the mean of the reported distribution changes as risk aversion increases as ρ increases.

Table 1: Experiment Design and Sample Sizes

Session	Treatments		Total
	10bin	**2bin**	
1	15	14	29
2	15	16	31
3	15	17	32
4	13	14	27
5	15		15
6	18		18
7	18		18
8	14		14
Total	123	61	184

Notes: Treatment **10bin** is elicitation of a *distribution* with the QSR, treatment **2bin** is elicitation of a *probability* with the QSR.

Table 2: Interval Regression Results

Session	True % of Red Balls (True)	Estimated Mean in **10bin** (μ_{10})	Standard Error of Estimated μ_{10}	95% Confidence Interval for μ_{10}		p-value for H_0: μ_{10} = True
				Lower Limit	Upper Limit	
			A. Without Controls			
1	0.69	0.71	0.03	0.66	0.77	0.42
2	0.13	0.14	0.03	0.08	0.21	0.70
3	0.10	0.12	0.006	0.11	0.13	0.002
4	0.57	0.59	0.04	0.51	0.66	0.67
5	0.41	0.37	0.02	0.32	0.42	0.11
6	0.62	0.59	0.02	0.55	0.64	0.27
7	0.11	0.25	0.06	0.12	0.37	0.03
8	0.33	0.35	0.03	0.30	0.41	0.42
			B. With Controls for Individual Heterogeneity of Variance			
3	0.10	0.11	0.007	0.098	0.12	0.09
7	0.11	0.15	0.03	0.097	0.20	0.13

Figure 4: Initial Belief Elicitation Interface

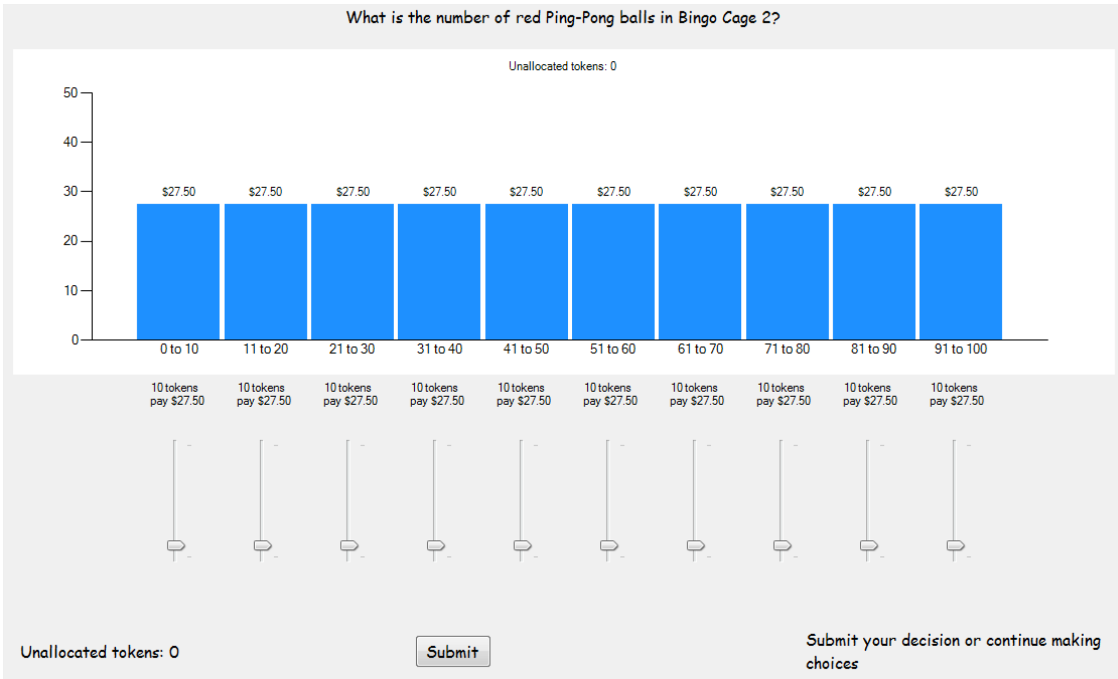


Figure 5: Typical Belief Elicitation Response

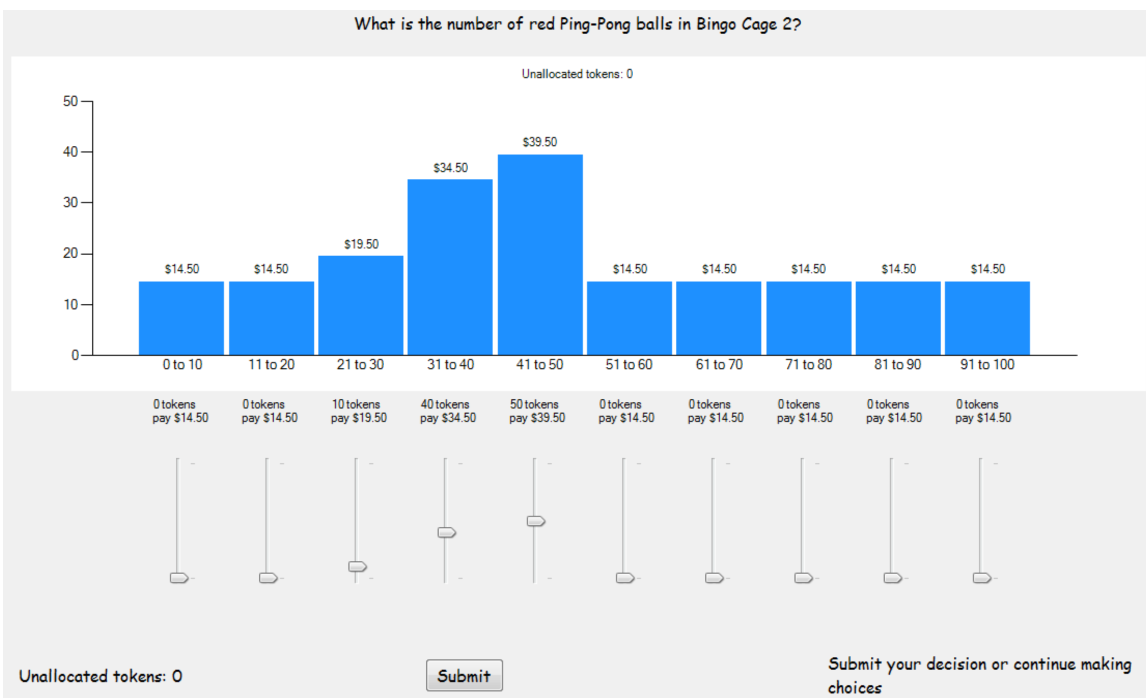


Figure 6: Elicited Subjective Distribution Pooled Over 15 Subjects and With True Percent of 69%

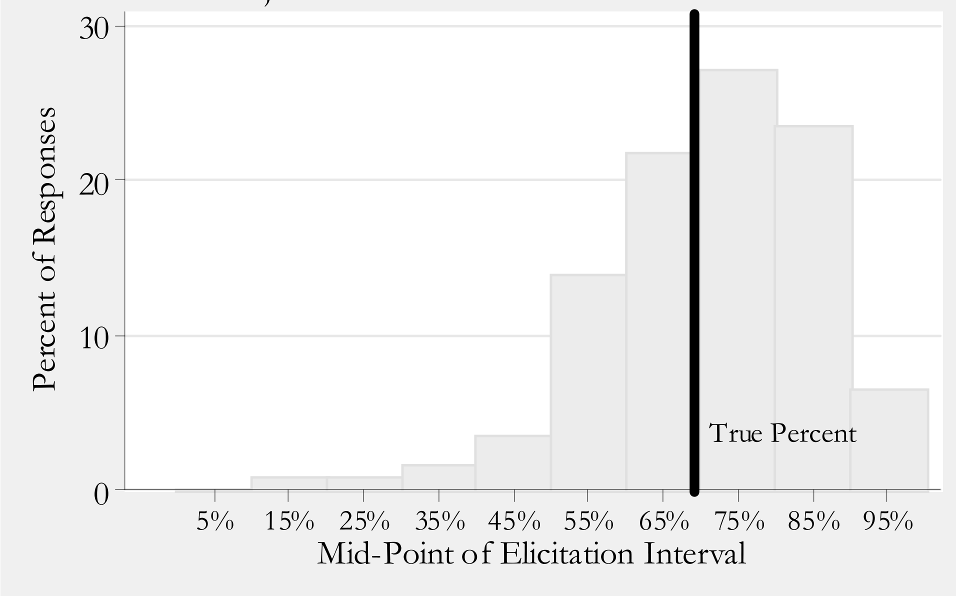


Figure 7: Average Elicited Subjective Belief Distribution

Pooled average for each of 8 sessions
Sample sizes: 15, 15, 15, 13, 15, 18, 18 and 14

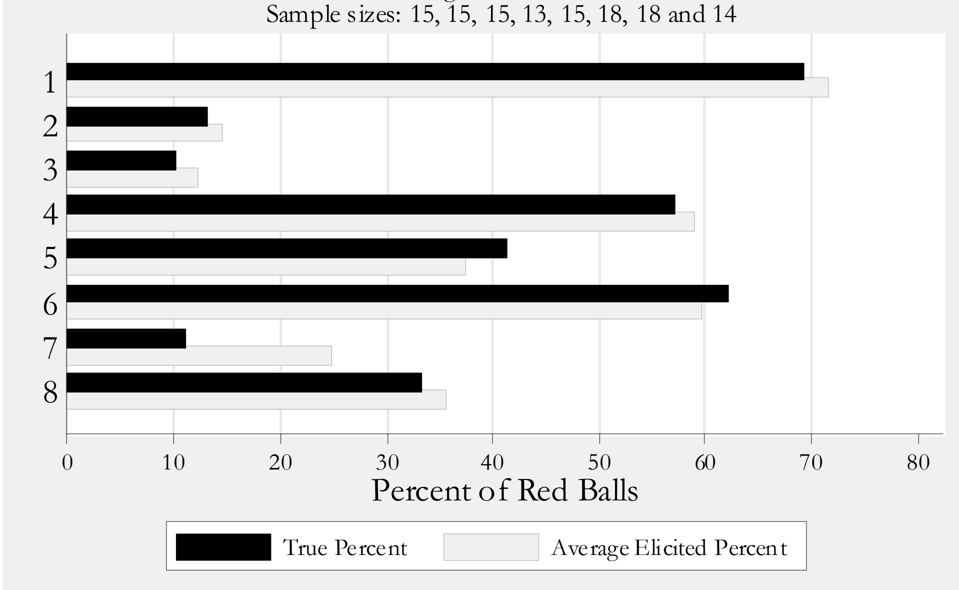


Figure 8: Subjective Probability Elicitation Interface

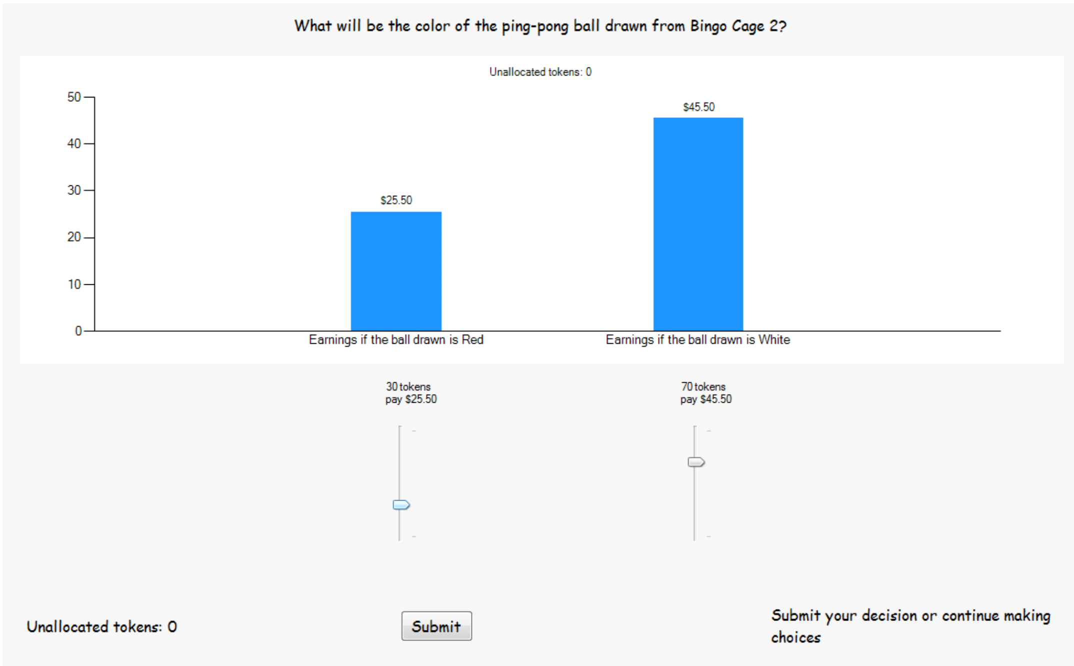
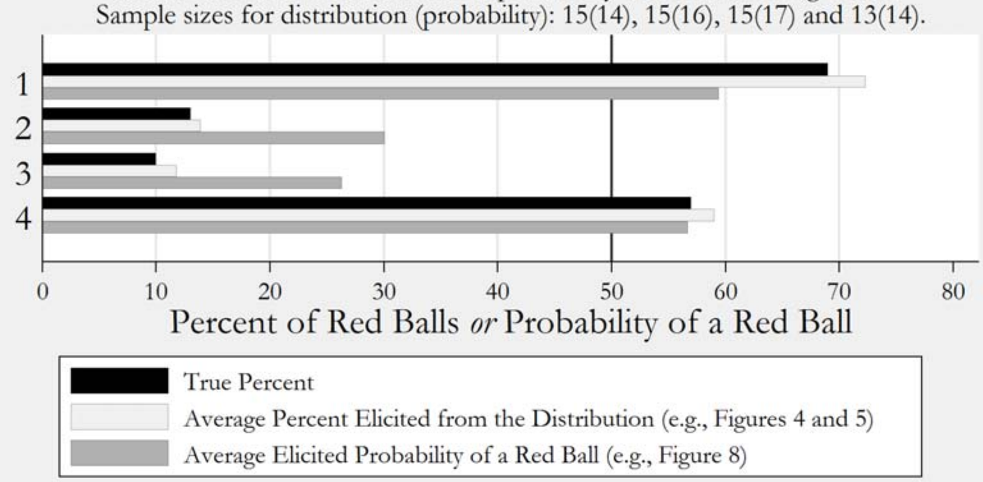


Figure 9: Average of Elicited Subjective Belief Distributions and Average Elicited Subjective Probabilities

Pooled averages for each of 4 sessions, with treatments *within* each session.
In each session, both treatments used the same random stimulus.
One treatment elicited beliefs about the true distribution of red balls,
and another treatment elicited the probability of a red ball being drawn.
Sample sizes for distribution (probability): 15(14), 15(16), 15(17) and 13(14).



References

Andersen, Steffen; Fountain, John; Harrison, Glenn W., and Rutström, E. Elisabet, "Estimating Subjective Probabilities," *Journal of Risk & Uncertainty*, 48, 2014, 207-229.

Antoniou, Constantinos; Harrison, Glenn W.; Lau, Morten I., and Read, Daniel, "Subjective Bayesian Beliefs," *Journal of Risk & Uncertainty*, 50, 2015, 35-54.

Antoniou, Constantinos; Harrison, Glenn W.; Lau, Morten I., and Read, Daniel, "Information Characteristics and Errors in Expectations: Experimental Evidence," *Journal of Financial & Quantitative Analysis*, 2016 forthcoming.

Armantier, Olivier and Treich, Nicolas, "Eliciting Beliefs: Proper Scoring Rules, Incentives, Stakes and Hedging," *European Economic Review*, 62, 2013, 17-40.

Costa-Gomes, Miguel A.; Huck, Steffen; Weizsäcker, Georg, "Beliefs and Actions in the Trust Game: Creating Instrumental Variables to Estimate the Causal Effect,"*Games and Economic Behavior*, 88, 2014, 298-309.

Cox, James C.; Roberson, Bruce; and Smith, Vernon L., "Theory and Behavior of Single Object Auctions," in V.L. Smith (ed.), *Research in Experimental Economics* (Greenwich: JAI Press, volume 2, 1982).

Delavande, Adeline; Giné, Xavier, and McKenzie, David, "Measuring Subjective Expectations in Developing Countries: A Critical Review and New Evidence," *Journal of Development Economics*, 94, 2011, 151-163.

Dominitz, Jeff, and Manski, Charles F., "Measuring and Interpreting Expectations of Equity Returns," *Journal of Applied Econometrics*, 26, 2011, 352-370.

Ergin, Haluk, and Gul, Faruk, "A Theory of Subjective Compound Lotteries," *Journal of Economic Theory*, 144, 2009, 899-929.

Fiore, Stephen M.; Harrison, Glenn W.; Hughes, Charles E., and Rutström, E. Elisabet, "Virtual Experiments and Environmental Policy," *Journal of Environmental Economics & Management*, 57, January 2009, 65-86.

Grether, David M., "Testing Bayes Rule and the Representativeness Heuristic: Some Experimental Evidence," *Journal of Economic Behavior & Organization*, 17, 1992, 31-57.

Harrison, Glenn W., "Hypothetical Surveys or Incentivized Scoring Rules for Eliciting Subjective Belief Distributions?" *Working Paper 2014-05*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2014.

Harrison, Glenn W., and Rutström, E. Elisabet, "Risk Aversion in the Laboratory," in J.C. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in

Experimental Economics, Volume 12, 2008).

Harrison, Glenn W., Martínez-Correa, Jimmy, and Swarthout, J. Todd, "Eliciting Subjective Probabilities with Binary Lotteries," *Journal of Economic Behavior & Organization*, 101, 2014, 128-140.

Harrison, Glenn W., Martínez-Correa, Jimmy, and Swarthout, J. Todd, "Reduction of Compound Lotteries with Objective Probabilities: Theory and Evidence," *Journal of Economic Behavior and Organization*, 119, 2015, 32-55.

Harrison, Glenn W., and Ng, Jia Min, "Evaluating the Expected Welfare Gain of Insurance," *Journal of Risk & Insurance*, 83, 2016, 91-120.

Harrison, Glenn W., and Phillips, Richard D., "Subjective Beliefs and the Statistical Forecasts of Financial Risks: the Chief Risk Officer Project," in T.J. Andersen (ed.) *Contemporary Challenges in Risk Management* (New York, Palgrave Macmillan, 2014).

Harrison, Glenn W., and Swarthout, J. Todd, "Experimental Payment Protocols and the Bipolar Behaviorist," *Theory and Decision*, 77, 2014, 423-438.

Harrison, Glenn W., and Ulm, Eric, "Recovering Subjective Belief Distributions," *CEAR Working Paper 2016-01*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2016.

Holt, Charles A., and Laury, Susan K., "Risk Aversion and Incentive Effects," *American Economic Review*, 92(5), 2002, 1644-1655.

Holt, Charles A., and Laury, Susan K., "Risk Aversion and Incentive Effects: New Data Without Order Effects," *American Economic Review*, 95(3), 2005, 902-912.

Holt, Charles A., and Smith, Angela M., "An Update on Bayesian Updating," *Journal of Economic Behavior & Organization*, 69, 2009, 125-134.

Kadane, J. B. and Winkler, Robert L., "Separating Probability Elicitation from Utilities," *Journal of the American Statistical Association*, 83, 1988, 357-363.

Karni, Edi, "A Mechanism for Eliciting Probabilities," *Econometrica*, 77, 2009, 603-606.

Karni, Edi, and Safra, Zvi, "The Impossibility of Experimental Elicitation of Subjective Probabilities," *Theory and Decision*, 38, 1995, 313-320.

Köszegi, Botond, and Rabin, Matthew, "Revealed Mistakes and Revealed Preferences," in A. Caplin and A. Schotter (eds.), *The Foundations of Positive and Normative Economics: A Handbook* (New York: Oxford University Press, 2008).

Lichtendahl, Jr., Kenneth C., and Winkler, Robert L., "Probability Elicitation, Scoring Rules, and

Competition Among Forecasters," *Management Science*, 53, 2007, 1745-1755.

Matheson, James E., and Winkler, Robert L, "Scoring Rules for Continuous Probability Distributions," *Management Science*, 22, 1976, 1087-1096.

Merkle, Christoph, and Weber, Martin, "True Overconfidence: The Inability of Rational Information Processing to Account for Apparent Overconfidence," *Organizational Behavior and Human Decision Processes*, 116, 2011, 262-271.

Moore, Don A., and Healy, Paul J., "The Trouble With Overconfidence," *Psychological Review*, 115, 2008, 502-517.

Nau, Robert F., "Uncertainty Aversion with Second-Order Utilities and Probabilities," *Management Science*, 52, 2006, 136-156.

Offerman, Theo; Sonnemans, Joep, and Schram, Arthur, "Value Orientations, Expectations and Voluntary Contributions in Public Goods," *Economic Journal*, 106, 1996, 817-845.

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

Savage, Leonard J., "Elicitation of Personal Probabilities and Expectations," *Journal of American Statistical Association*, 66, 1971, 783-801.

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

Appendix A: Proofs for the Discrete Case

Lemma 1. Suppose a subjective discrete probability distribution $\{p_1, p_2, ..., p_k, ..., p_K\}$ over K states of nature and utility function $u(w)$ over random wealth. If the subject is given a scoring rule determined by $w(k) = \alpha + \beta 2r_k - \beta \sum_{i=1...K} (r_i)^2$, then the optimal report $r = \{r_1, r_2, ..., r_k, ..., r_K\}$ solves the following problem:

$$\mathbf{Max}_{\{r\}}\ E_p[\, u(w)\,] \text{ subject to } \sum_{i=1...K} (r_i) = 1 \tag{A.2}$$

where $E_p[\, u(w)\,] = \sum_{j=1...K} p_j \times u[\, \alpha + \beta 2r_j - \beta \sum_{i=1...K} (r_i)^2\,]$. In some experimental configurations there may be K additional constraints: $r_i \geq 0$ for $i = 1,..., K$. These constraints are not included in (A.2) because they are automatically satisfied by the solution (A.1) for both risk-averse and risk-loving individuals.

Problem (A.2) can be solved by maximizing the Lagrangian

$$\mathcal{L} = \sum_{j=1...K} p_j \times u[\, \alpha + \beta 2r_j - \beta \sum_{i=1...K} (r_i)^2\,] - \lambda\, [\, \sum_{i=1...K} (r_i) - 1]. \tag{A.3}$$

The solution to the problem must satisfy K+1 conditions. The K first order conditions with respect to report r_k, $\forall\, k = 1,..., K$, are

$$\partial\mathcal{L}/\partial r_k = \sum_{j=1...K} (\, p_j \times \partial u(w(j))/\partial r_k\,) - \lambda = 0, \forall\, k = 1,..., K \tag{A.4}$$

where $\partial u(w(j))/\partial r_k = \partial u/\partial w\,|_{w=w(j)} \times (2\beta\delta_{jk} - 2\beta \times r_k)$ and δ_{jk} is equal to 1 if $j = k$ and equal to zero if $j \neq k$. The (K+1)-*th* condition is the first order derivative of (A.3) with respect to the Lagrangian constant

$$\sum_{i=1...K} (r_i) - 1 = 0. \tag{A.5}$$

We can simplify the K equations in (A.4) as

$$2\beta p_k \times (\, \partial u/\partial w\,|_{w=w(k)}\,) - 2\beta r_k \sum_{j=1...K} p_j \times (\, \partial u/\partial w\,|_{w=w(j)}\,) - \lambda = 0, \forall\, k = 1,..., K$$

or

$$p_k \times (\, \partial u/\partial w\,|_{w=w(k)}\,) - r_k E_p\, [\, \partial u/\partial w\,] = \lambda/2\beta, \forall\, k = 1,..., K. \tag{A.4′}$$

Summing over the K first-order conditions we get

$$E_p\, [\, \partial u/\partial w\,|_{w=w(k)}\,] - \sum_{k=1...K} r_k E_p\, [\, \partial u/\partial w\,] = K\lambda/2\beta\,. \tag{A.6}$$

Notice that $\sum_{k=1...K} r_k E_p\, [\, \partial u/\partial w\,] = E_p\, [\, \partial u/\partial w\,]$ because the expectation term is a constant and because of (A.5). Then (A.6) implies that $K\lambda/2\beta = 0$, which can only be satisfied if $\lambda = 0$ since $K>0$ and $\beta>0$. This result and (A.4′) implies that the solution to problem (A.2) must satisfy the following K conditions:

$$p_k \times \partial u/\partial w\,|_{w=w(k)} - r_k \times E_p[\partial u/\partial w] = 0, \forall\, k = 1,..., K. \qquad \blacksquare$$

Lemma 2. Assume that the conditions of Lemma 1 in (A.1) are satisfied and the distortions between the actual and reported probabilities are given by $r_k = p_k + \varepsilon_k$, with $\sum_{k=1...K} \varepsilon_k = 0$. Define $f_k = \partial u/\partial w \mid_{w=w(k)}$ and $f = \partial u/\partial w$. Then the K conditions in (A.1) become

$$p_k \times f_k - p_k \times E_p[f] - \varepsilon_k \times E_p[f] = 0, \forall k = 1,..., K. \quad (A.1')$$

Solving for ε_k we get the K conditions stated in Lemma 2:

$$\varepsilon_k = p_k \times \{f_k - E_p[f]\}/E_p[f], \forall k = 1,..., K. \quad ■$$

Lemma 3. The intuition for the result is as follows. Suppose that $\{r_1, r_2, ..., r_k, ..., r_K\}^*$ is a solution to (A.2). Then if $p_i = p_j$ for some i and j, the subject must assign the same weight to reports in states i and j, that is $r_i = r_j$. The proof is then by contradiction.

Assume that $p_i = p_j$ for some i and j. Now suppose without loss of generality that $r_i > r_j$. By definition of the deviation of subjective and reported probabilities the latter implies that $\varepsilon_i > \varepsilon_j$ because

$$r_i = p_i + \varepsilon_i > r_j = p_j + \varepsilon_j. \quad (A.8)$$

Since $r_i > r_j$, we also know that $w(i) > w(j)$, and by the concavity of $u(.)$ the latter implies that $f_i < f_j$. Therefore

$$\{f_i - E_p[f]\}/E_p[f] < \{f_j - E_p[f]\}/E_p[f]. \quad (A.9)$$

But by Lemma 2, (A.9) implies that $r_i < r_j$, which is a contradiction. ■

Lemma 3 does not hold for risk loving individuals. The following counterexample proves it. Suppose that $u(w) = w^2$, $p_1 = ½$ and $p_2 = ½$, $\alpha = 0$ and $\beta = 1$. We see that

$$\begin{aligned} E_p[\, u(w(r_1))\,] &= 0.5\,(2\,r_1 - r_1^2 - (1\text{-}r_1)^2\,)^2 + 0.5\,(2\,(1\text{-}r_1) - r_1^2 - (1\text{-}r_1)^2\,)^2 \\ &= 4\,r_1^4 - 8\,r_1^3 + 8\,r_1^2 - 4\,r_1 + 1 \\ \partial E_p[\,u\,]/\partial r_1 &= 16\,r_1^3 - 24\,r_1^2 + 16\,r_1 - 4. \end{aligned}$$

To maximize subjective EU set the first order condition equal to zero, and then check the end points $r_1 = 0$ and $r_1 = 1$. We then have

$$\begin{aligned} r_1^3 - (1½)\,r_1^2 + r_1 - ¼ &= 0 \\ (r_1 - ½)(r_1^2 - r_1 + ½) &= 0. \end{aligned}$$

Solving for the real root we get $r_1 = ½$. By reporting $r_1 = ½$, the subjective EU is equal to ¼, while if the report is $r_1 = 1$ or $r_1 = 0$ the subjective EU is equal to 1. Thus symmetry is broken, and the optimal report is ($r_1 = 1$, $r_2 = 0$) or ($r_1 = 0$, $r_2 = 1$): that is, $p_1 = p_2$ but $r_1 \neq r_2$.

Proposition 1. A symmetric subjective distribution for random variable y with mean μ is one of two types: odd and even. Take the case of the odd type first. Consider a subjective probability

p_k and report r_k, for $k = 1, ..., n$, with n being an odd integer. Let $m = (n+1)/2$ such that the subjective probability p_m is the likelihood that the random variable takes the value of μ. Also let p_{m-i} and p_{m+i} be, respectively, the subjective probability that the random variable takes the value of $\mu-\eta_i$ and $\mu+\eta_i$, for $i = 1...m-1$ and $p_{m-i} = p_{m+i}$.

By Lemma 3, $r_{m-i} = r_{m+i}$,

$$E_p[\,y\,] = \Sigma_{i=1...m-1}\, p_{m-i}\,(\mu-\eta_i) + \Sigma_{i=1...m-1}\, p_{m+i}\,(\mu+\eta_i) + p_m\,\mu \qquad (A.10)$$
$$= \Sigma_{j=1...n}\, p_j\,\mu + \Sigma_{i=1...m-1}\,[\,p_{m-i} - p_{m+i}\,]\,\eta_i = \mu + 0 = \mu,$$

and

$$E_r[\,y\,] = \Sigma_{i=1...m-1}\, r_{m-i}\,(\mu-\eta_i) + \Sigma_{i=1...m-1}\, r_{m+i}\,(\mu+\eta_i) + r_m\,\mu \qquad (A.11)$$
$$= \Sigma_{j=1...n}\, r_j\,\mu + \Sigma_{i=1...m-1}\,[\,r_{m-i} - r_{m+i}\,]\,\eta_i = \mu + 0 = \mu.$$

By (A.10) and (A.11) we have that $E_p[y]-E_r[y] = 0$. The even case is similar except that $m = n/2$ and the random variable taking a value equal to μ has no weight. ■

Lemma 4. Follows from Lemma 1. If $r_i = r_j$ then

$$p_i\,\partial u/\partial w\,|_{r_i} - r_i\,E_p[\partial u/\partial w]=0 \text{ and } p_j\,\partial u/\partial w\,|_{r_j} - r_j\,E_p[\partial u/\partial w]=0. \qquad (A.12)$$

Thus,

$$p_i = r_i\,E_p[\partial u/\partial w]/\,\partial u/\partial w\,|_{r_i} = r_j\,E_p[\partial u/\partial w]/\,\partial u/\partial w\,|_{r_j} = p_j\,. \qquad ■$$

Proposition 2. Identical to Proposition 1, with r_k and p_k, $\forall k$, interchanged at all steps. ■

Proposition 3. Using Lemma 2, if $p_i = 0$, then $\varepsilon_i=0$ and $r_i = 0$. The converse claim follows from Lemma 1: since $\partial u/\partial w\,|_{w=w(k)}$ and $E_p[f]$ are both positive, if $r_k = 0$ then $p_k = 0$. ■

Proposition 4. We will show that $\exists\, p^*$ such that if $p_k > p^*$ then $p_k > r_k > p^*$ and p^* is the value such that

$$\partial u/\partial w\,|_{w=w(p^*)} = E_p[\,\partial u/\partial w\,].$$

From Lemma 2 we know that

$$\varepsilon_k = p_k \times \{\partial u/\partial w\,|_{w=w(k)} - E_p[\,\partial u/\partial w\,]\}/E_p[\,\partial u/\partial w\,],\ \forall\, k=1,..., K.$$

We also know that $w(k)$ is monotonically increasing in r_k , and therefore $\partial u/\partial w$ is monotonically decreasing in r_k . If $r_k > (<)\ p^*$, $\varepsilon_k<(>)0$, $r_k < (>)\ p_k$ by definition. Since $p^* <(>)\ r_k$, then $p^* <(>)\ r_k <(>)\ p_k$. ■

Proposition 5. By Lemma 2 we know that $r_k = p_k \times \{\partial u/\partial w\,|_{w=w(k)}\}/E_p[\,\partial u/\partial w\,]$. Let p^* be selected such that $\partial u/\partial w\,|_{w=w(p^*)} = E_p[\,\partial u/\partial w\,]$. Then let $u(w) = w - c\,[w - w(p^*)]^2$ without loss of generality. Therefore

$$E_p[\,\partial u/\partial w\,]=\partial u/\partial w\,|_{w=w(p^*)} =1.$$

Let $r_k = p^* + \delta_k$ be the deviations in reports with respect to p^* due to risk aversion. Additionally,

$$w(k) = \alpha + \beta 2 r_k - \beta \sum_{i=1...K} (r_i)^2$$
$$w(k) = \alpha + \beta 2(p^* + \delta_k) - \beta \sum_{i \neq k} [(p^* + \delta_i)^2] - \beta (p^* + \delta_k)^2 \quad \text{(A.13)}$$

and

$$w(p^*) = \alpha + \beta 2 p^* - \beta \sum_{i=1...K} (r_i)^2$$
$$w(p^*) = \alpha + \beta 2(p^*) - \beta \sum_{i \neq k} [(p^* + \delta_i)^2] - \beta (p^* + \delta_k)^2. \quad \text{(A.14)}$$

Both (A.13) and (A.14) imply that $w(k) - w(p^*) = \beta 2 \delta_k$. Taking the derivative of the utility function with respect to w and evaluating at $w(k)$, we obtain

$$\partial u / \partial w \,|_{w=w(k)} = 1 - 2c\,[w(k) - w(p^*)] = 1 - 2c\,[\,\beta 2 \delta_k] = 1 - 4c\,\beta \delta_k. \quad \text{(A.15)}$$

By the definition of r_k, $\partial u / \partial w \,|_{w=w(k)}$ and $E_p[\,\partial u / \partial w\,]$ we have

$$r_k = p^* + \delta_k = p_k \times \{\partial u / \partial w \,|_{w=w(k)}\} / E_p[\,\partial u / \partial w\,],$$

which implies that

$$p^* + \delta_k = p_k \times \{1 - 2c\,[\,\beta 2 \delta_k]\,\} / \{1\}.$$

Solving for δ_k we obtain $\delta_k = \{p_k - p^*\} / \{1 + 4c\,\beta\, p_k\,\}$. If $p_k \neq 0$, then $\lim_{c \to \infty} \delta_k = 0$ and the deviations become vanishingly small for sufficiently risk-averse individuals.

Now prove that $p^* \approx 1/K$, where K is the number of states for which $p_k \neq 0$. By definition

$$\sum_{i=1...K} (r_i) = \sum_{i=1...K} (p^* + \delta_k).$$

If $p_k = 0$, then $\lim_{c \to 0} \delta_k = p_k - p^*$ and $\lim_{c \to \infty} r_k = p^* + \delta_k = p_k = 0$. If $p_k \neq 0$, then $\sum_{p^k \neq 0} (\delta_k)$ tends to zero and $\sum_{p^k \neq 0} (p^*) = 1 = K p^* = 1$, so $p^* = 1/K$ in the limit. These two facts combine to prove that if $p_k \neq 0$ then $\lim_{c \to \infty} r_k = \lim_{c \to \infty} p^* + \delta_k = 1/K$. That is, the reported probabilities approach a uniform distribution over the outcomes where the subjective probability is non-zero. ■

Proposition 6. If a subject exhibits utility function $u(w) = w + \delta \times u^*(w)$, we know from (A.1′) that the following K conditions must be satisfied:

$$p_k \times [1 + \delta \times \partial u^* / \partial w \,|_{w=w(k)}] - p_k \times \{1 + \delta\, E_p[\,\partial u^* / \partial w\,]\} - \boldsymbol{\varepsilon}_k \times \{1 + \delta\, E_p[\,\partial u^* / \partial w\,]\} = 0,\ \forall\ k = 1,..., K,$$

where $\boldsymbol{\varepsilon}_k$ is defined in (A.7) for Lemma 2. Solving for $\boldsymbol{\varepsilon}_k$ we obtain

$$\boldsymbol{\varepsilon}_k = \delta\, p_k \times \{\partial u^* / \partial w \,|_{w=w(k)} - E_p[\,\partial u^* / \partial w\,]\} / \{1 + \delta\, E_p[\,\partial u^* / \partial w\,]\},\ \forall\ k = 1,..., K. \quad \text{(A.16)}$$

Assume a random variable y with K possible states of nature. Define $E_r[\,y\,] = \sum_{k=1...n} r_k\, y_k$ and $E_p[\,y\,] = \sum_{k=1...n} p_k\, y_k$. Then the difference of the expected value of y under measures $\{r_1, r_2, ..., r_k, ..., r_K\}$ and $\{p_1, p_2, ..., p_k, ..., p_K\}$ is equal to $E_r[\,y\,] - E_p[\,y\,] = \sum_{k=1...K} \boldsymbol{\varepsilon}_k\, y_k$. Substituting for $\boldsymbol{\varepsilon}_k$ using (A.16), it can be shown that the denominator $\{1 + \delta\, E_p[\,\partial u^* / \partial w\,]\}$ drops out (take a Taylor Series expansion of the reciprocal, multiply terms with the numerator, and drop higher-order terms). Then we have

$$E_r[\,y\,] - E_p[\,y\,] \approx \delta \times \sum_{k=1...K} p_k\{\partial u^*/\partial w\,|_{w=w(k)} - E_p[\,\partial u^*/\partial w\,]\}\; y_k$$
$$\approx \delta \times \{E_p\,[\,\partial u^*/\partial w \times y\,] - E_p\,[\,\partial u^*/\partial w\,]\;\; E_p\,[\,y\,]\}$$
$$\approx \delta \times Cov_p[\partial u^*/\partial w, y] = Cov_p[\partial u/\partial w, y]. \qquad \blacksquare$$

Proposition 7. From Proposition 6 we know that

$$E_r[\,y\,] - E_p[\,y\,] \approx \delta \times \sum_{k=1...K} p_k\{\partial u^*/\partial w\,|_{w=w(k)} - E_p[\,\partial u^*/\partial w\,]\}\; y_k \qquad \text{(A.17)}$$

Now, $p_k = r_k - \boldsymbol{\varepsilon}_k$ and

$$\boldsymbol{\varepsilon}_k = \delta \times p_k \times \{\partial u^*/\partial w\,|_{w=w(k)} - E_p[\,\partial u^*/\partial w\,]\},$$

also from proposition 6. Therefore,

$$E_r[\,y\,] - E_p[\,y\,] \approx \delta \times \sum_{\text{k=1...K}} r_k\{\partial u^*/\partial w\,|_{w=w(k)} - E_p[\,\partial u^*/\partial w\,]\}\; y_k$$
$$- \delta^2 \times \sum_{\text{k=1...K}} p_k\{\partial u^*/\partial w\,|_{w=w(k)} - E_p[\,\partial u^*/\partial w\,]\}^2\; y_k. \qquad \text{(A.18)}$$

The second term above is a second order in δ and can be removed. Thus

$$E_r[\,y\,] - E_p[\,y\,] \approx \delta \times \sum_{\text{k=1...K}} r_k\{\partial u^*/\partial w\,|_{w=w(k)} - E_p[\,\partial u^*/\partial w\,]\}\; y_k. \qquad \text{(A.19)}$$

We now see that

$$E_p[\,\partial u^*/\partial w\,] \approx \sum_{k=1...K} p_k\, \partial u^*/\partial w\,|_{w=w(k)}$$
$$\approx \sum_{k=1...K} r_k\, \partial u^*/\partial w\,|_{w=w(k)} - \delta \sum_{k=1...K} p_k\, \{\partial u^*/\partial w\,|_{w=w(k)} - E_p[\,\partial u^*/\partial w\,]\}\partial u^*/\partial w\,|_{w=w(k)}$$
$$\approx E_r[\partial u^*/\partial w\,] - \delta\,[E_p[\,(\partial u^*/\partial w\,)^2] - (E_p[\,(\partial u^*/\partial w\,)])^2]$$
$$\approx \mathrm{E_r}[\partial u^*/\partial w\,] - \delta\, Var_p[\partial u^*/\partial w\,].$$

Substituting into (A.19) we get

$$E_r[\,y\,] - E_p[\,y\,] \approx \delta \times \sum_{\text{k=1...K}} r_k\{\partial u^*/\partial w\,|_{w=w(k)} - E_r[\,\partial u^*/\partial w\,]\}\; y_k$$
$$+ \delta^2 \times \sum_{\text{k=1...K}} r_k\; Var_p[\partial u^*/\partial w\,]\;\; y_k. \qquad \text{(A.20)}$$

Eliminating the second order term gives us

$$E_r[\,y\,] - E_p[\,y\,] \approx \delta \times \sum_{\text{k=1...K}} r_k\{\partial u^*/\partial w\,|_{w=w(k)} - E_r[\,\partial u^*/\partial w\,]\}\; y_k$$
$$\approx \delta \times \{E_r[\,\partial u^*/\partial w \times y] - E_r[\,\partial u^*/\partial w\,] \times E_r[\,y\,]\}$$
$$\approx \delta \times Cov_r[\partial u^*/\partial w, y]$$
$$\approx Cov_r[\partial u/\partial w, y] \qquad \blacksquare$$

To prove Proposition 8 we must first prove Theorem 1, which is interesting in its own right. Lemma 1 will then follow for all proper scoring rules. Since Proposition 8 states that Propositions 1-7 generalize and those propositions follow from Lemma 1, Proposition 8 will be proved. We follow Armantier and Treich [2013] who proved the result for 2 elicitation bins. We prove an analogous theorem for an arbitrary number of bins.

Define a scoring rule S where $S_1(r_1,..., r_n)$, $S_2(r_1,..., r_n)$,..., and $S_n(r_1,..., r_n)$ represent the payoffs for each of the possible states of nature 1,..., n. S_k is the payoff if state k is realized after reports r_1,..., r_n, where $r_n = 1 - \sum_{i=1...n-1} r_i$. Let

$$f(p_1,..., p_n; r_1,..., r_n) = \sum_{i=1...n} p_i S_i(r_1,..., r_n).$$

A scoring rule is "proper" if the maximizing arguments are $r_i = p_i$ for all i.

Theorem 1: A scoring rule is proper if and only if there exists a function $g(q_1,..., q_{n-1})$ with conditions on the second derivatives guaranteeing uniqueness and maximization such that

$$S_n(q_1,..., q_{n-1}) = g - \sum_{j=1...n-1} q_j \, \partial g/\partial q_j$$

and
$$S_j(q_1,..., q_{n-1}) = S_n(q_1,..., q_{n-1}) + \partial g/\partial q_j \text{ for } j \in [1,n-1].$$

Notice that q_n is not an argument in the functions anymore because the latter is defined by q_1,..., q_{n-1}.

<u>Necessity (only if)</u>.

Let $g(q_1,..., q_{n-1}) = \max_{\{r^*\}} f(q_1,..., q_{n-1}; r_1,..., r_{n-1})$ where $r^* = \{r_1^*, r_2^*,..., r_{n-1}^*\}$ is the vector of reports that maximizes the function f.

By the envelope theorem, we see that

$$\partial g/\partial q_j = \partial f(q_1,..., q_{n-1}; r_1,..., r_{n-1})/\partial q_j|_{r^i = q^i \forall i}$$

$$= S_j(q_1,..., q_{n-1}) - S_n(q_1,..., q_{n-1}).$$

Notice that $S_n(q_1,..., q_{n-1})$ comes from a $(1 - \sum_{i=1...n-1} r_i) \, S_n(r_1,..., r_{n-1})$ term. Therefore

$$S_j(q_1,..., q_{n-1}) = S_n(q_1,..., q_{n-1}) + \partial g/\partial q_j.$$

Substituting these into the formula for g, we get

$$g(q_1,..., q_{n-1}) = \max_{\{r^*\}} f(q_1,..., q_{n-1}; r_1,..., r_{n-1}) = f(q_1,..., q_{n-1}; q_1,..., q_{n-1}),$$

since S is a proper scoring rule.

Therefore,

$$g(q_1,..., q_{n-1}) = \sum_{j=1...n-1} q_j [S_n(q_1,..., q_{n-1}) + \partial g/\partial q_j] + (1 - \sum_{j=1...n-1} q_j) \, S_n(q_1,..., q_{n-1})$$

$$= S_n(q_1,..., q_{n-1}) + \sum_{j=1...n-1} q_j \, \partial g/\partial q_j.$$

Rearranging terms we get

$$S_n\,(q_1,..., q_{n-1}) = g(q_1,..., q_{n-1}) - \Sigma_{j=1...n-1}\, q_j\, \partial g/\partial q_j.$$

Sufficiency (if).

$$f\,(q_1,..., q_{n-1}; r_1,..., r_{n-1}) = \Sigma_{i=1...n-1}\, q_i\, S_i\,(r_1,..., r_{n-1}) + (1 - \Sigma_{i=1...n-1}\, q_i)\, S_n\,(r_1,..., r_{n-1})$$

$$= \Sigma_{i=1...n-1}\, q_i\, [g - \Sigma_{j=1...n-1}\, r_j\, \partial g/\partial r_j + \partial g/\partial r_i]$$

$$+ (1 - \Sigma_{i=1...n-1}\, q_i)\, (g - \Sigma_{j=1...n-1}\, r_j\, \partial g/\partial r_j)$$

We maximize f by setting the n-1 first order conditions to zero:

$$\partial f/\partial r_k = \Sigma_{i=1...n-1}\, q_i\, [\partial g/\partial r_k - \Sigma_{j=1...n-1}\, r_j\, \partial^2 g/\partial r_j\, \partial r_k - \partial g/\partial r_k + \partial^2 g/\partial r_i\, \partial r_k]$$

$$+ (1 - \Sigma_{i=1...n-1}\, q_i)\, (\partial g/\partial r_k - \Sigma_{j=1...n-1}\, r_j\, \partial^2 g/\partial r_j\, \partial r_k - \partial g/\partial r_k) = 0.$$

This gives us

$$- \Sigma_{i=1...n-1}\, q_i\, \Sigma_{j=1...n-1}\, r_j\, \partial^2 g/\partial r_j\, \partial r_k + \Sigma_{i=1...n-1}\, q_i\, \partial^2 g/\partial r_i\, \partial r_k - \Sigma_{j=1...n-1}\, r_j\, \partial^2 g/\partial r_j\, \partial r_k$$

$$+ \Sigma_{i=1...n-1}\, q_i\, \Sigma_{j=1...n-1}\, r_j\, \partial^2 g/\partial r_j\, \partial r_k = 0.$$

Cancelling terms, we obtain

$$\Sigma_{i=1...n-1}\, q_i\, \partial^2 g/\partial r_i\, \partial r_k - \Sigma_{j=1...n-1}\, r_j\, \partial^2 g/\partial r_j\, \partial r_k = 0.$$

Changing the index from j to i in the second summation of the first order condition above we see that

$$\Sigma_{i=1...n-1}\, (q_i - r_i)\, \partial^2 g/\partial r_i\, \partial r_k = 0. \qquad \text{(A.21)}$$

This system consists of n-1 equations (indexed by k) in the n-1 unknowns $(q_i - r_i)$ indexed by i . One solution is clearly $q_i - r_i = 0$ (or $q_i = r_i$) for all i. Thus, the scoring rule S is *proper*.

There must be conditions on the second derivatives of g such that this solution is *unique* and *maximizes*, rather than minimizes, f. ■

Now we can prove Lemma 1 for general proper scoring rules.

Suppose an individual is now trying to maximize utility $V(p_1,..., p_{n-1}; r_1,..., r_{n-1})$ rather than money $f(p_1,..., p_{n-1}; r_1,..., r_{n-1})$. Suppose a utility function of wealth $u(W)$. We have

$$V(p_1,..., p_{n-1}; r_1,..., r_{n-1}) = \Sigma_{j=1...n-1}\, p_j\, u(S_j\,(p_1,..., p_{n-1})) + p_n\, u(S_n\,(p_1,..., p_{n-1})), \text{ where } \Sigma_{j=1...n}\, p_j = 1.$$

We solve the following n-1 first order conditions to maximize:

$$\partial V/\partial r_k = \Sigma_{j=1...n-1}\, p_j\, \partial u/\partial W|_{s_j}\, \partial S_j\, /\partial r_k + p_n\, \partial u/\partial W|_{s_n}\, \partial S_n\, /\partial r_k.$$

Now, since $S_j = S_n + \partial g/\partial r_j$, we see

$$\partial S_j/\partial r_k = \partial S_n/\partial r_k + \partial^2 g/\partial r_j \partial r_k$$

and $$\partial V/\partial r_k = \Sigma_{j=1...n}\, p_j\, \partial u/\partial W|_{s_j}\, \partial S_n/\partial r_k + \Sigma_{j=1...n-1}\, p_j\, \partial u/\partial W|_{s_j}\, \partial^2 g\, /\partial r_j\, \partial r_k = 0$$

$$= \partial S_n/\partial r_k\, \Sigma_{j=1...n}\, p_j\, \partial u/\partial W|_{s_j} + \Sigma_{j=1...n-1}\, p_j\, \partial u/\partial W|_{s_j} \partial^2 g\, /\partial r_j\, \partial r_k = 0$$

$$= \partial S_n/\partial r_k\, E_p[\partial u/\partial W] + \Sigma_{j=1...n-1}\, p_j\, \partial u/\partial W|_{s_j}\, \partial^2 g\, /\partial r_j\, \partial r_k = 0$$

where $E_p[\ .\ ]$ denotes the expectations operator under probability measure $p = \{p_1,...,p_n\}$. Now, since $S_n = g - \Sigma_{j=1...n-1}\, r_j\, \partial g/\partial r_j$, we get

$$\partial S_n/\partial r_k = \partial g/\partial r_k - \Sigma_{j=1...n-1}\, r_j\, \partial^2 g\, /\partial r_j\, \partial r_k - \partial g/\partial r_k = - \Sigma_{j=1...n-1}\, r_j\, \partial^2 g\, /\partial r_j\, \partial r_k,$$

so $$\partial V/\partial r_k = - \Sigma_{j=1...n-1}\, r_j\, \partial^2 g\, /\partial r_j\, \partial r_k\, E_p[\partial u/\partial W] + \Sigma_{j=1...n-1}\, p_j\, \partial u/\partial W|_{s_j}\, \partial^2 g\, /\partial r_j\, \partial r_k = 0.$$

Therefore, we obtain

$$\Sigma_{j=1...n-1}\, [\, p_j\, \partial u/\partial W|_{s_j} - r_j\, E_p[\partial u/\partial W]]\, \partial^2 g\, /\partial r_j\, \partial r_k = 0. \qquad \text{(A.22)}$$

Equation (A.22) looks just like equation (A.21) except the n-1 unknowns are

$$p_j\, \partial u/\partial W|_{s_j} - r_j\, E_p[\partial u/\partial W].$$

As before,

$$p_j\, \partial u/\partial W|_{s_j} - r_j\, E_p[\partial u/\partial W] = 0\ \forall\ j.$$

This is unique and maximizing from the convexity conditions on g. ■

Since Propositions 1-7 follow from Lemma 1, Proposition 8 has been proved.