Revealed Preference and the Strength/Weight Hypothesis

Constantinos Antoniou, Glenn W. Harrison, Morten I. Lau, Daniel Read

April 2013

Abstract

The strength/weight hypothesis is the claim that people are more persuaded by extreme evidence (high strength) than by reliable evidence (high weight), even when these two dimensions of evidence are diagnostically equivalent according to Bayes Rule. We investigate whether this hypothesis is supported when tested in a revealed preference experiment. We confirm the hypothesis. Although earlier researchers using hypothetical, stated preference methods have found a pattern of overconfidence for high strength/low weight information, and underconfidence for low strength/high weight information, we find that underconfidence is the main direction of the bias. Moreover, the strength/weight effect in our revealed preference experiment was much smaller than in earlier stated preference studies.

† Xfi Centre for Finance and Investment, Business School, University of Exeter, U.K. (Antoniou); Department of Risk Management & Insurance and CEAR, Robinson College of Business, Georgia State University, USA (Harrison); Durham University Business School, UK (Lau); and Behavioural Science Group, Warwick Business School, University of Warwick (Read). Email: c.antoniou@exeter.ac.uk, gharrison@gsu.edu, m.i.lau@durham.ac.uk and Daniel.Read@wbs.ac.uk.

1. Introduction

The diagnostic value of information depends on both its extremity and its reliability. As an illustration, consider two earnings forecasts for different companies and by different analysts. Analyst S predicts an extreme increase in earnings-per-share (EPS) of 15% for his company, which he has been following for a year. Analyst W predicts a moderate increase of 3% in EPS for her company, but she is an experienced and reliable analyst who has been following her company for more than a decade. Imagine an investor who receives both forecasts. Which company should this investor expect to have the greater increase in earnings? The answer depends on the appropriate rate at which information extremity and information reliability should be traded off. Griffin and Tversky (1992) refer to information extremity as strength, and to information reliability as weight. A strong forecast with low weight (Analyst S) might warrant a weaker, equal or stronger response than a moderate forecast with high weight (Analyst W). Bayes rule provides a procedure for making this trade-off. We investigate how well people conform to this Bayesian ideal.

Griffin and Tversky (1992) hypothesized what we call the *strength/weight effect*, which is that if the Bayesian diagnosticity of information is held constant, judgment and choice will be more influenced by strength than weight. In our example, if a Bayesian analysis would indicate that both forecasts should lead their respective recipients to draw the same conclusions about the EPS of the two firms, it is nonetheless likely that the expectation formed by the recipient of the extreme but less reliable forecast will be viewed as more favorable.

The strength/weight hypothesis has been highly influential, especially in economics where it has been used to interpret phenomena that apparently contradict the efficient market hypothesis. Barberis, Shleifer and Vishny (1998), for example, argued that the strength/weight effect explains momentum and reversals in stock prices (Jegadeesh and Titman, 1993; DeBondt and Thaler, 1985). According to Barberis et al. (1998), momentum occurs when the market underreacts to low strength/high weight signals, while price reversals can occur when the market overreacts to high strength/low weight signals. Massey and Wu (2005) proposed that the effect can account for the failure of organizations to recognize “regime shifts,” or situations in which the underlying market fundamentals have changed. Related applications were proposed by Sorescu and Subrahmanyam (2006), Liang (2003), Bloomfield and Hales (2002) and Nelson et al. (2001).

We test the strength/weight effect experimentally using a novel experimental procedure, drawing on the same conceptual logic as Griffin and Tversky (1992), who operationalized strength as the proportion of a sample showing an attribute and weight as the size of that sample. One major departure from Griffin and Tversky (1992) is the use of methods drawn from experimental economics, including the use of a physical process to generate samples varying in strength and weight, and the reliance on revealed rather than stated preference. Griffin and Tversky (1992) tested the strength/weight effect by asking people to give verbal probability judgments. We tested the effect by observing choices and inferring the underlying probabilities that are consistent with these choices, obviating the need for explicit and truthful reporting (or even awareness) of subjective probabilities. Despite this change in design we confirm the strength/weight effect, although it is substantially reduced relative to that reported by Griffin and Tversky (1992).

Our design also allows us to investigate further questions. One is whether the strength/weight effect would be reflected in how confidently people hold their beliefs, or in the "precision" of subjective probabilities. Griffin and Tversky (1992) initially undertook to reconcile two apparently contradictory findings in decision science, overconfidence and underconfidence (or "conservatism"). In their studies they took the stated probability of the focal hypothesis as a measure of confidence, and the signed difference between the stated probability and the correct Bayesian posterior as a measure of overconfidence. The strength/weight effect can thus be restated as "people are relatively more confident when they are responding to high strength/low weight than to low strength/high weight information."

In addition to stated probabilities, confidence is also widely measured as belief *precision*, usually indexed by eliciting a range of values within which which the decision maker feels the correct value might fall (a psychological confidence interval). Confidence about a stated probability would therefore be a stated range that the decision maker believes has an X% chance of containing the correct probability. In our experiment we measured belief precision as the variation in choice probabilities emerging from repeated encounters with the same information. This alternative measure of confidence reveals a somewhat weaker strength/weight effect. Another innovation of our study is to examine how inferences about the strength/weight effect change when we allow for risk aversion. We jointly estimate risk aversion coefficients and choice probabilities using a structural maximum likelihood model. To estimate risk aversion we use choices made over a series of lotteries in a separate task, following Hey and Orme (1994).

The next section discusses the relevant literature and gives an overview of our contribution. The third section describes our experimental procedures. The fourth section presents the results when we assume risk neutrality. The fifth section presents analysis that allows for risk aversion, and the sixth section summarizes and discusses our findings.

2. Literature and experimental overview

Griffin and Tversky's (1992) test of the strength/weight effect used a simple experimental setting, which we followed in its essentials. Their respondents judged the probability a coin would be biased in a particular direction based on the outcome of samples of hypothetical coin spins that varied in strength and weight. Respondents were first asked to imagine there were two classes of coins that differed in their bias to fall heads up or tails up when spun, with one class coming up heads 60% of the time and the other coming up heads 40% of the time. Respondent then estimated the probability the coin being spun was a "60%" coin based on sample evidence consisting of *N* coin spins, where the sample size *N* and the number of heads in that sample was varied systematically.

Griffin and Tversky (1992) operationalized the weight of the evidence as the size of each sample of coin spins (*N*), and strength as the proportion of heads in that sample, or *h/N*, where *h* is the numbers of heads.[1] A sample of 3 heads and 0 tails, for example, has Weight = 3 and Strength =1, while a sample of 10 heads and 7 tails has Weight = 17 and Strength = 10/17. Both samples, however, have equal diagnosticity, with Bayes' rule giving a posterior of 0.77. Nonetheless, and consistent with the strength/weight hypothesis, Griffin and Tversky's (1992) respondents gave higher hypothetical stated probabilities for the first high strength/low weight sample than for the second low strength/high weight sample -- the average probability estimates were 0.85 and 0.60, respectively. In general, the hypothetical stated probability for the high strength/low weight samples were too high (overconfidence), and those for the Bayesian-equivalent low strength/high weight sample were too low (underconfidence or "conservatism"). A regression analysis confirmed that probability judgments increased more rapidly with Bayesian-equivalent increases in strength than weight.

Kraemer and Weber (2004) examined the strength/weight effect by also eliciting hypothetical stated probabilities, using as respondents students of economics and decision-making. In their experiment respondents were asked to imagine there were two possible states of the world, either A or B, and that they had received a series of "*a*" and "*b*" signals which varied in "quality." Quality was interpreted for respondents as follows: "a signal that has a quality of 60% can be interpreted as a random draw from an urn containing 60% 'correct' signals, and 40% 'wrong' signals (drawn with replacement)" (p. 137). In fact, all of the experimental signals in the experiment had a quality of 60%, meaning there was a 60% chance that an "*a*" signal would be observed if the state of the world was A.

The participants in Kraemer and Weber's (2004) study were informed of hypothetical signal sets, having sample sizes of either 5, 15 or 25. In every set the number of *a* and *b* signals differed by only 1. For instance, when the sample size was 25 there were 13 *a* and 12 *b* signals. The correct posterior probability was, therefore, always 0.6. For each signal set, respondents provided a stated-probability by means of a slider bar. Respondents were incentivized by being paid in inverse proportion to how much their stated probabilities diverged from the correct Bayesian posterior. This formal procedure differed from Griffin and Tversky (1992), who paid $20 to the respondent whose judgments "most closely" matched the correct values.

Our design advances on those used by both Kraemer and Weber (2004) and Griffin and Tversky (1992). First, as already mentioned, we measure choice-probabilities, revealed in the incentivized decisions made by participants, rather than hypothetical stated-probabilities. Our method is based on the principles of subjective probability elicitation outlined by Ramsey (1931) and Savage (1954; 1971). Our respondents chose between chose between bets, or options varying in the payoff they offered if different states of the world were true, and we inferred the underlying subjective probabilities from the pattern of bets made.

Such choice probabilities have advantages over stated probabilities. First, they better reflect the fact that the key economic role of information is as a basis for action, and not for talk. Second, they do not depend on the ability to translate feelings or judgments be translated into verbal reports, or indeed that there even be any feelings or judgments to translate. Ramsey himself, in 1931, had expressed doubts about the possibility of making probability statements, and underlined the primacy of the propensity to act in the measurement of probability.

> … when we seek to know what is the difference between believing more firmly and believing less firmly, we can no longer regard it as consisting in having more or less of certain observable feelings; at least *I personally cannot recognize any such feelings. The difference seems to me to lie in how far we should act on these beliefs:* this may depend on the degree of some feeling or feelings, but I do not know exactly what feelings and I do not see that it is indispensable that we should know. (p. 170-171, italics added)[2]

Doubts about whether people have privileged introspective access to cognitive processes and valuations has increased since Ramsey's time, with the growth of a vast and still increasing body of evidence that key drivers of behavior are inaccessible to consciousness and therefore *cannot* be reported (see Pronin, 2009; Wilson, 2003). The process of stating probabilities need not access, as Ramsey (1931) assumed, "what determines us to act," nor does it need to even reflect those determinants.[3] And this is quite apart from the lack of any incentive to report accurately.

To extend this argument, it is possible that if people do not have access to a "feeling" of probability, then stated probabilities may be based on what they are *able* to report. To illustrate one possibility, consider that the sample strength in Griffin and Tversky (1992) is the proportion of successes in a sample (*h/N*). This proportion is easy to compute, and if respondents simply used this "proportion heuristic" to answer the question "what is the probability this is a 60% coin?" then the strength/weight effect would be perfectly confirmed (Berman, 2007).[4] In fact, the median probabilities reported by Griffin and Tversky (1992) are highly consistent with this interpretation, since they are very close to those ratios. This general procedure, in which an easy judgment (a ratio) is substituted for a hard one (a Bayesian posterior) has been termed "attribute substitution" by Kahneman and Frederick (2002).

A further difference between our experiment and earlier studies is that the samples in previous studies were hypothetical outcomes from hypothetical processes. That is, in the earlier studies respondents were asked to imagine a given outcome was the result of either spinning a coin (Griffin and Tversky, 1982) or an unspecified signal generation process (Kraemer & Weber, 2004). Both earlier studies restricted what respondents saw to a subset of samples that favored one specific hypothesis, and even within that restriction certain samples could never occur. We generated random samples by means of a physical generating process that actually occurred in front of our respondents. Moreover, respondents saw, and responded to, all samples drawn using that physical process.

A final innovation is that we were able to examine how inferences about the magnitude of the strength/weight effect are influenced when we drop the assumption of risk-neutrality. We elicited preferences toward risk using a series of lotteries based on Hey and Orme (1994), and use these measures to adjust subjective choice-probability estimates for non-linearity in the utility function. Kraemer and Weber (2004) implicitly used a linear scoring rule to incentivize respondents, and this scoring rule also needs to be corrected for the distorting effect of risk aversion (Andersen, Fountain, Harrison and Rutström (2010)).

3. Methods

3.1 Participants

We recruited 111 respondents from the University of Durham, UK. All received a £5 show up fee. Payments for the experiment totalled £2,692, for an average payment of £24.26 per subject. Table 1 gives sample demographics.

3.2 Experimental procedures

The experimental session included two tasks: belief elicitation, in which choices were made that allowed us to infer choice probabilities, and the measurement of risk attitude. The two tasks were counterbalanced. Respondents participated in groups of about 10. Our focus is on the belief elicitation task, although we use the risk attitude task to control for the effects of non-linear utility.

We implemented the Savage/Ramsey method of probability elicitation. There were two equally likely states of the world, the "White" and the "Blue." Respondents were provided with relevant sample information, after which they chose between pairs of acts (or "bets") that offered different payoffs depending on which state of the world actually obtained. Choice probabilities were inferred from the pattern of acts chosen.

The belief elicitation task was based around a physical randomization device, as in other studies investigating Bayesian inference in the laboratory (Grether, 1980; Holt and Smith, 2009). The randomization devices were 6-sided and 10-sided dice. Part of the design was implemented in private, so one participant in each group was chosen as a "monitor" who supervised the rolling and counting of dice, announced the outcomes, and kept the experimenters honest in the minds of the respondents. The monitor was chosen randomly from amongst all the participants in each session, and received a flat payment of £10 for the session.

There were three cups with dice. The *Choice* cup contained one 6-sided die with three blue and three white sides, and was used to select between the *Blue* and the *White* cup. Both colored cups contained *N* 10-sided dice, where *N* was either 3, 5, 9 and 17. These values were chosen to match the number of coin spins in Griffin and Tversky (1992; Experiment 1). The White cup had *N* dice with six white and four blue sides, and the Blue cup had *N* dice with six blue and four white sides. The die in the Choice cup was rolled first. If blue came up we rolled the dice in the Blue cup, otherwise we rolled all the dice in the White cup. This procedure was conducted behind a screen but observed by the monitor. We then announced the outcome of the *N* random draws and wrote the information on a whiteboard visible to all.

After the sample information was announced, respondents placed "bets" on White and Blue, using a decision sheet adapted from Fiore, Harrison, Hughes and Rutström (2009). They were asked to conceptualize the task as one of making 19 separate bets. These bets were presented simultaneously in a multiple price list. Each row of the list offered a choice between betting on the White or the Blue set, with the odds of each outcome differing across the 19 rows. Figure 1 shows the betting interface used to elicit subjective beliefs about posterior probabilities. The complete set of instructions for the belief elicitation task is reproduced in Appendix 2. To illustrate how inferences are made from this task, imagine a subject who chose to bet on White in row 5 and Blue in row 6. If the subject was risk neutral, the bet on White for row 5 would indicate a probability of White higher than 25%, while the bet on Blue for row 6 would indicate a probability of White of 30% or less. This subject's choice probability would therefore be bracketed between 25% and 30%.

In each session respondents saw 30 samples, 4 samples of three dice (i.e., N=3), 14 of five dice, 6 of nine dice and 6 of seventeen dice. The distribution of sample sizes was chosen to ensure that, across sample sizes, we roughly equalized the frequency of the least likely sample distributions.[5] Once responses to all 30 samples were completed we randomly picked one sample to be played out for each subject and then randomly picked one of the 19 bets that the subject made for that sample. The subject was then paid according to the outcome of that bet.

As explained by Savage (1971) and Kadane and Winkler (1988), subjective probabilities estimated on the assumption of risk neutrality will not correspond to "true" latent subjective probabilities if the respondents are not risk neutral. The intuition is obvious from inspection of the two possible earnings in Figure 1: risk averse bettors are drawn to report closer to the bet that gives them £6 in each state of the world.[6] We controlled for risk attitudes using the experimental design of Hey and Orme (1994). All respondents made a series of 20 choices between two lotteries, each offering between 1 and 3 prizes valued at £5, £10 or £15. Figure 2 displays a typical lottery pair. The 20 choices presented to each respondent comprised one of three subsets of 20, drawn from a superset of 60 choices. At the end of the task one of the 20 choices made by each subject was selected at random and played out for money. Appendix 3 reproduces the instructions for this task.

3.3 Testing the Strength/Weight hypothesis

To test the strength/weight hypothesis we must first separate strength from weight, and then test how much each component predicts choice probabilities. Recall that respondents made decisions based on which of two sets of dice had been rolled. Each set had N dice with blue and white faces, with the dice in one set having a 60% chance of a blue face; and those in the other set having a 40% chance of a blue face (or 60% chance of white). The strength of the evidence is the ratio of successes to trials, while its weight is the sample size N.

As discussed in Appendix 1, for computational purposes, and following earlier researchers, we normalize strength to facilitate the interpretation of regression analyses. This normalized index, which we denote S, is the absolute value of the difference in the number of successes for each outcome, divided by N. If we denote the number of blue and white faces in a sample as b and w, then $S = abs(b-w)/N$. Appendix 1 shows that in this Bayesian framework the log odds ratio for the two mutually exclusive hypotheses can be written as:

$$log\ \{log(\pi/(1-\pi))\ /\ log(0.6/0.4)\ \} = \alpha\ log\ N + \beta\ log(S) \qquad (1)$$

We estimate (1) with the following statistical model, setting the left-hand side of the equation equal to y for simplicity:

$$y_{i,j} = c + \alpha \log N_j + \beta \log(S)_j + e_{i,j}, \quad (2)$$

where $y_{i,j}$ is the log odds ratio implied by the choices made by individual *i* in task *j*. If Bayes' Rule is followed without error, then $\alpha=\beta=1$ and $c=0$. That is, a given change in the log of the sample size should have the same impact as the same change in the log of *S*. The strength/weight hypothesis is that $\alpha<\beta$, or that a marginal increase in strength will increase choice-probabilities more than a marginal increase in weight. The strength/weight hypothesis implies no restriction on *c*.

4. Results assuming risk neutrality

4.1 Choice probability

The analysis in this section assumes that respondents are risk neutral, an assumption that is relaxed in Section 5. Table 2 presents the main findings. The first column shows the Bayesian posterior. This posterior, and all statistics in the table, are for the most likely set. That is, for samples having more white than blue sides coming up (Panel A) it is the probability of the White set, otherwise when more blue than white sides coming up (Panel B) it is the probability of the Blue set. The second and third column presents the weight and the strength ratio. The fourth column shows the mean estimates of choice probabilities under the assumption of risk neutrality using an interval regression model, and the standard errors of these estimates are in the fifth column. The sixth and seventh columns show the estimated posterior probabilities and standard errors respectively when we controlled for risk aversion (Section 5). The final column shows the frequency of each sample.[7]

One way of stating the strength/weight hypothesis is that if we hold the posterior probability constant, then choice-probabilities will increase in the strength of the evidence (or, equivalently, decrease in weight). This hypothesis is strongly supported, as can be seen by scanning down the values for choice-probabilities from Table 2 within each Bayesian probability group. For each group, mean choice-probability increases in strength.

Table 3 depicts results from testing the model in (2) using a truncated regression model.[8] Because we have multiple observations from the same individual (i.e., $j=1,\ldots,n$ for each i), the standard errors were calculated by clustering on the subject level (a procedure used for all statistical analyses). Consistent with the strength/weight hypothesis, we found that strength has roughly twice the influence on subjective probabilities as weight; the estimated coefficient of strength is 1.92 with a standard error of 0.52, and the estimated coefficient of weight is 0.94 with a standard error of 0.28.[9] The hypothesis that the strength and weight coefficients are equal is safely rejected ($p < 0.001$). The constant term in the regression is also reliably different from 0 ($p < 0.001$), which suggests there is some "systematic noise" in subjective probabilities. Overall these results support the strength/weight hypothesis.

Figure 3 shows a comparison between our findings and those of Griffin and Tversky (1992). We plotted choice probabilities for each pattern in each posterior probability group against the stated-probabilities from Griffin and Tversky (1992, Table 1 p.415). The first column in each of the three posterior groups (0.60, 0.77 and 0.88) shows the median stated probabilities obtained by Griffin and Tversky (1992) for signals of different strength. The second column depicts the *median* choice probabilities (not the means from Table 2) from our data when we assume risk neutrality, and the third column depicts estimated choice probabilities when we allow for risk aversion. The dotted line in each group shows the correct Bayesian posterior.

The comparison shows that, although the strength/weight effect is strongly present in both datasets, it is more pronounced in Griffin and Tversky (1992). For the posterior group with Bayesian probability equal to 0.88, for instance, they report a bias of roughly 28%, while in our study the bias is 15%. This reduction could be due to a number of reasons, as the two experimental designs are very different, with the reduced opportunity to use the proportion heuristic being a likely contributor.[10]

An additional result is that Griffin and Tversky (1992) found a pattern of overconfidence for high strength samples, and underconfidence for low strength samples. We found the general leaning was toward underconfidence, with approximately correct levels of confidence for high-strength samples.

4.2 Strength, Weight and Belief Precision

Our random sampling procedure meant each subject typically made more than one response for formally identical samples. For instance, a subject might make several choices based on the observation of three white and two blue sides for the roll of five dice. We were therefore able to measure how "confidently" the choice probabilities were held, by looking at the variation in those probabilities from trial to trial.[11] We calculate the variance of subjective probabilities at all strength/weight combinations and investigate the expanded strength/weight hypothesis, which is that higher strength not only leads to higher choice-probabilities, but also to choice-probabilities that are less likely to vary from estimate to estimate.

Whenever the same respondent saw an identical sample *n* times, we obtained *n* choice probabilities from their responses. We calculate the standard deviation of these probabilities, σ_{ψ}, as a measure of the dispersion in the subjective distribution of probabilities that underlies the specific sample, and then estimate a statistical model relating it to strength and weight. Because the distribution of σ_{ψ} is highly non-normal (it assumes only positive values since it is a standard deviation, and 40% of the observations are equal to 0) we used a maximum likelihood hurdle specification.[12] These results show whether relative increases of strength decrease the dispersion in this subjective distribution of probabilities, in effect making individuals more confident.

The hurdle model has two parts. The first is a probit specification where we model the relationship between the explanatory variables and the probability that the dependent variable assumes a value greater than a threshold, taken here to be 0. In this specification a negative coefficient on strength, for example, implies that increases in strength reduce the probability that the dependent variable is greater than 0. Behaviourally this then implies that increases in strength reduce the dispersion in the subjective distribution of probabilities, in effect making respondents more confident. The second specification is a constrained OLS model where we model the relationship between strength, weight and the standard deviation, σ_ψ, conditional on σ_ψ exceeding the 0 threshold. Again a negative coefficient implies lower dispersion and therefore greater confidence.

The results are shown in Table 4. The probit model in Panel A shows that the probability of σ_ψ being larger than 0 decreases significantly with both strength (coefficient: -0.727, p <0.001) and weight (coefficient: -0.042, p <0.001), which suggests that increases in these variables increase confidence. From the constrained OLS model in Panel B we observe that strength continues to exert a negative impact on σ_ψ (coefficient: -0.359, p=0.025), whereas the effect of *Weight* is insignificant (coefficient: -0.010, p=0.174). Due to the non-linearity of the estimator in Table 4 we also provide the marginal effects associated with strength and weight in both the probit and the constrained OLS specifications.

Overall, the evidence from this test provides support to the strength/weight hypothesis. Whereas both strength and weight appear to have an equal effect on the probability of respondents having non-zero dispersion in their beliefs, strength is negatively related to the magnitude of this dispersion, whereas weight is not.

4.3 The "cost" of the strength/weight heuristic

Using the strength/weight heuristic will incur economic costs. In this section we provide an estimate for these costs and compare them with the costs implied by the magnitude of biases reported by Griffin and Tversky (1992). For each specific dice outcome we examine how a representative subject, whose beliefs are given by the median choice probabilities depicted in Figure 3, would place their bets, and how those bets would differ from those made by a Bayesian subject. We also consider the expected cost to this representative subject for being non-Bayesian.

The analysis is summarized in Table 5. The first two columns describe the dice rolls in terms of strength and weight. The next columns show the Bayesian probability, the representative median probabilities from our study (OS), and from Griffin and Tversky (GT). The next columns show the decisions that would be made by the representative subject when facing the odds in our multiple price list (collapsing over the symmetric halves of the price list) The dark line dividing the table in two shows the switching point of a risk neutral Bayesian respondent, and the labels OS and GT indicate where, respectively, the OS agent and the GT agent would choose differently than the Bayesian agent. If these different choices are to the left of the Bayesian line it indicates underconfidence, if they are to the right it indicates overconfidence. The final two columns show the expected loss in earnings, in percentage terms, for the OS and GT agents relative to the Bayesian one.

The economic cost for an agent behaving consistently with our estimates is very low. Over all samples, the average loss is just over 2% of earnings, if we assume all dice rolls are equally likely. This is substantially below the cost of an agent who behaves consistently with the estimates of Griffin and Tversky (1992), which is almost 6%. All the cost in our study is due to underconfidence, or switching too soon.

5. Subjective Expected Utility preferences

Choice probabilities will be affected by risk attitude. To illustrate why, go back to our earlier illustration of an agent who bets on White in row 5 and Blue in row 6. If the agent was risk neutral, this would imply a choice-probability interval for Blue between 70% and 75% (or equivalently 25% and 30% for the White). Now imagine a risk averse agent, in the context of Expected Utility Theory (EUT), whose risk preferences are characterized by a Constant Relative Risk Aversion (CRRA) utility function with r = 0.5, assuming a functional form of the type $u(x) = x^{1-r}/(1-r)$. For the risk-neutral agent we computed the relevant choice-probabilities by directly taking the ratio between the White payoff and the sum of the White and Blue payoffs. When we allow for non-linear utility we adjust by taking the ratio between the *utility* of the White payoff, and the sum of the utilities of the two payoffs. For bookie #5 this would be: U(£4)/(U(£4)+U(£1.33)) = 63%, and for bookie #6 this would be U(£3.33)/(U(£3.33)+U(£1.43)) = 60%. Therefore, the inferred probability range for the Blue-set changes from {70%, 75%} to {60%, 63%}.

Figure 4 shows this more systematically. It depicts the objective probability implied by different switch points. However, these switch points do not reflect the latent belief of a risk averse individual. Due to diminishing marginal utility this agent requires a larger payoff to bet on the alternative that is least likely. So he will *delay* his switching point until that payoff is reached. This implies that his belief is closer to 0.5 than the observed switch point actually suggests. The y-axis in Figure 4 shows the implied, risk averse probability for an agent with CRRA utility and different values of r. The figure shows that risk neutral probabilities that are furthest from 0.5 undergo the largest transformation, and that this transformation is increasing in r.[13]

We can use the choices made in the risk task to construct a likelihood function to estimate risk attitude, assuming CRRA and EUT. We use the choices made in the belief task to construct another likelihood from which we estimate choice probabilities, conditional on the risk attitude estimated from the risk task. The resulting parameters (risk attitudes and choice probabilities) are those that jointly maximize the sum of the two likelihood functions. The joint estimation approach is discussed in more detail in Appendix 4.

The coefficient of risk aversion we estimate from this joint estimation is equal to 0.562 and is highly significant, indicating risk aversion. The sixth column in Table 2 shows the estimate of the choice probabilities for each dice pattern obtained from the joint estimation. While the gap between the high and low strength choice probabilities of equal posterior is reduced, it remains strongly monotonic in the direction predicted by Griffin and Tversky (1992).

We can estimate the model in (1), controlling for risk attitude by modifying the log likelihood according to equations (10)-(13) in Appendix 4 and estimating the coefficients on strength and weight. The results are shown in Table 7. We obtain a larger coefficient on strength than on weight (0.736 vs. 0.442). The hypothesis that these coefficients are equal is safely rejected. However, the difference between the two coefficients is smaller compared to the risk neutral case.

6. Conclusion

6. 1 Summary

Most decisions are made in the face of uncertainty due to incomplete information. Griffin and Tversky (1992) and others (Kraemer and Weber 2004; Wu and Massey, 2005) have developed the strength/weight hypothesis, which is that decision makers are more responsive to the extremity (strength) of the information they have than to its predictive validity (weight), even when both strength and weight are equally diagnostic. We re-visit the strength/weight hypothesis by means of an experiment using a novel design that allowed people to reveal their subjective "choice probabilities" through their gambling decisions over real monetary rewards. We provide respondents with imperfect information about the true state of the world, and ask them to *reveal* their subjective belief about the likelihood of the true state by making a series of bets. The design is based on the logic of Savage (1954; 1971) and Ramsey (1926).

Our results confirmed the predictions of Griffin and Tversky (1992). We found that, holding objective probability constant, decision makers perceived events as more likely when the available evidence had high strength and low weight, than when it had low strength and high weight. Although this pattern matched the one observed by Griffin and Tversky (1992), in a qualitative sense there were some differences. First, the effect we observed was considerably smaller. The associated monetary costs of the bias we observed was less than half of that which would be incurred based on the responses elicited by Griffin and Tversky (1992). Second, whereas Griffin and Tversky (1992) observed a pattern of overconfidence (or too-high probability estimation) for high strength/low weight samples we observed either no bias for these samples or slight conservatism.

6.2 Discussion

The strength/weight effect appears to be robust, although less quantitatively important than previously found, but what is its underlying cause? One possibility is that people are generally less able to incorporate some features of decision-relevant information (weight) into their decision than others (strength). Another possibility is that the effect arises from how specific information is made more or less salient by the context in which it is presented. We suspect the strength/weight effect is driven by both forces. Weight is harder to incorporate into decision making than strength, but how much harder depends on the choice context. In the case of strength and weight, there is a fundamental asymmetry between them since strength can be applied with little contextual support, while weight requires much more of this support.

To illustrate how strength is generally easier than weight to incorporate into decision making we can return to the example with which we started. Imagine you are considering the purchase of a stock. You have limited access to a single analyst, about whom you know nothing. The analyst could be very reliable or very unreliable – he or she is a random draw from the distribution of all advisors. You can choose one piece of information – either how reliable the analyst is, or the forecast that analyst will make about your stock. Which do you choose? It is clear that while the forecast value (or strength) provides some information even if you don't know the forecaster's reliability (weight), the forecaster's reliability provides no information if you don't know the forecast. Similarly, in the context of our experiment, imagine two people each given a piece of information about a sample of dice rolls. Person A is told only that 17 dice were rolled (weight only), while Person B is told that approximately 60% of the dice rolled showed a blue face (strength only). Again only Person B can use what they have learned to help them make a decision. One reason strength has primacy in our thinking might be that it is useful by itself, while weight is only interpretable when combined with strength.

Building on this point, it is likely that even if both strength and weight information are available, and even if we do not inherently undervalue weight, we will give primacy to strength because of this ready interpretability. Our attention must be expressly drawn to weight information and its significance to give it its due "weight." An example of apparent weight neglect is given in investors' response to advertising by mutual funds (see Jain & Wu, 2000; Koehler & Mercer, 2009). Koehler and Mercer (2009) observed that mutual fund companies typically advertise only a high performing subset of funds, and not their entire fund portfolio. In our language, the strength of a company's advertisement would be the performance of the advertised funds, while one component of weight would be the number of funds in the company's portfolio that are not advertised. Koehler and

Mercer (2009) found that people were completely insensitive to weight information, treating a company that advertised all of its funds just like one who advertised only a subset of funds, unless their attention was explicitly drawn to this fact. This finding illustrates again that evidence on the reliability of information typically captures too little attention unless it is explicitly highlighted and perhaps even interpreted for the respondent.

References

Andersen, S., Fountain, J., Harrison G. W., and Rutström, E., "Estimating Subjective Probabilities," *Working Paper 2010-06*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2010

Antoniou, C., Harrison G.W., Lau, M. and Read, D., "Subjective Bayesian Beliefs," *Working Paper 2012-08*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2012.

Barberis, N. and Shleifer, A., and Vishny, R., "A Model of Investor Sentiment," *Journal of Financial Economics*. 49, 307-343, 1998.

Berman, J., "On the Misuse of Using Bayesian Inference Tasks to Infer Cognitive Biases," *Master's Thesis,* London School of Economics, 2007.

Bloomfield, R. and Hales, J., "Predicting the Next Step of a Random Walk: Experimental Evidence of Regime Shifting Beliefs," *Journal of Financial Economics,* 65, 397-414, 2002.

Costa-Gomes M. A., and Weizsäcker, G., "Stated Beliefs and Play in Normal-Form Games," *Review of Economic Studies*, 75, 729-762, 2008.

De Bondt, W. F. M., and Thaler, R., "Does the Stock Market Overreact," *Journal of Finance,* 40, 793-805, 1985.

Engelberg, J., Manski, C. F., and Williams, J., "Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters," *Journal of Business & Economic Statistics,* 27, 30-41, 2009.

Fiore S.M., Harrison G.W., Hughes, C.E., and Rutström E. E., "Virtual Experiments and Environmental Policy," *Journal of Environmental Economics and Management,* 57, 65-86, 2009.

Gilboa I., Postlewaite A. W., and Schmeidler D., "Probability and Uncertainty in Economic Modeling," *Journal of Economic Perspectives,* 22, 173-188.

Grether D.M., "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *Quarterly Journal of Economics*, 95, 537-557, 1980.

Grether D. M., and Plott C. R., "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review,* 69, 623-638, 1979.

Griffin, D., and Tversky A., "The Weighing of Evidence and the Determinants of Confidence," *Cognitive Psychology*, 24, 411-435, 1992.

Harrison, G.W., and Rutström E. E., "Risk Aversion in the Laboratory," in J.C. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments,* Bingley, U.K: Emerald, Research in Experimental Economics, Vol. 12, 41-197, 2008.

Hey, J.D., and Orme C., "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica,* 62, 1291-1326, 1994.

Holt, C. A., *Markets, Games and Strategic Behavior,* New York: Addison-Wesley, 2006.

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

Jain C. P. and Wu, S., J., "Truth in Mutual Fund Advertising: Evidence on Future Performance and Fund Flows," *Journal of Finance,* 55, 937-958, 2000.

Jegadeesh, N., and Titman, S., "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance*, 48, 65-91, 1993.

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

Kahneman, D., and Frederick, S., "Representativeness Revisited: Attribute Substitution in Intuitive Judgment," In Thomas Gilovich, Dale Griffin and Daniel Kahneman (eds.). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press, 2002.

Koehler, J. J. and Mercer, M., "Selection neglect in mutual fund advertisements," *Management Science, 55*, 1107-1121, 2009.

Kraemer, C., and Weber, M., "How Do People Take into Account Weight, Strength and Quality of Segregated Vs. Aggregated Data? Experimental Evidence," *Journal of Risk and Uncertainty*, 29, 113-142, 2004.

Liang, L., "Post-Earnings Announcement Drift and Market Participants' Information Biases," *Review of Accounting Studies*, 8, 321-345, 2003.

Massey, C. and Wu, G., "Detecting Regime Shifts: The Causes of Over and Underreaction," *Management Science*, 51, 923-957, 2005.

Moore, D. A. and Healy, P. J., "The trouble with overconfidence," *Psychological Review,* 115, 502-517, 2008.

Nelson, M.W., Bloomfield R., Halles J., W. and Libby, R., "The Effect of Information Strength and Weight on Behavior in Financial Markets," *Organizational Behavior and Human Decision Processes*, 86, 168-196, 2001.

Plott, C. R., and Zeiler, K., "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, 2005.

Pronin, E. (2009). The introspection illusion. In Mark P. Zanna, editor: *Advances in Experimental Social Psychology*, Vol. 41, p. 1-67, Burlington: Academic Press.

Ramsey, F. P., "Truth and Probability," in F. P. Ramsey, *The Foundations of Mathematics and other Logical Essays*, Ch. VII, p.156-198, edited by R.B. Braithwaite, London: Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace and Company, 1931.

Rutström, E.E. and Wilcox N.T., "Stated Beliefs versus Inferred Beliefs: A Methodological Inquiry and Experimental Test," *Games and Economic Behavior*, 67, 616-632, 2009.

Savage, L., J., *The Foundations of Statistics,* New York: Dover, 1972; Second Edition.

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

Sorescu, S., and Subrahmanyam, A., "The Cross Section of Analyst Recommendations," *Journal of Financial and Quantitative Analysis*, 41, 139-168, 2006.

Wilson, T. D., "Knowing when to ask: Introspection and the adaptive unconscious," *Journal of Consciousness Studies*, 10, 131-140, 2003.

Zarnowitz, V., and Lambros, L., A., "Consensus and Uncertainty in Economic Prediction," *Journal of Political Economy*, 95, 591-621, 1987.

Figure 1: The Betting Sheet

Make a bet with all bookies.

Bookie	Stake	Odds offered		Earnings including the stake of £3		I will bet on (circle)
		White	Blue	White	Blue	
1	£3	20.00	1.05	£60.00	£3.15	W B
2	£3	10.00	1.11	£30.00	£3.33	W B
3	£3	6.67	1.18	£20.00	£3.54	W B
4	£3	5.00	1.25	£15.00	£3.75	W B
5	£3	4.00	1.33	£12.00	£4.00	W B
6	£3	3.33	1.43	£10.00	£4.29	W B
7	£3	2.86	1.54	£8.58	£4.62	W B
8	£3	2.50	1.67	£7.50	£5.00	W B
9	£3	2.22	1.82	£6.66	£5.46	W B
10	£3	2.00	2.00	£6.00	£6.00	W B
11	£3	1.82	2.22	£5.46	£6.66	W B
12	£3	1.67	2.50	£5.00	£7.50	W B
13	£3	1.54	2.86	£4.62	£8.58	W B
14	£3	1.43	3.33	£4.29	£10.00	W B
15	£3	1.33	4.00	£4.00	£12.00	W B
16	£3	1.25	5.00	£3.75	£15.00	W B
17	£3	1.18	6.67	£3.54	£20.00	W B
18	£3	1.11	10.00	£3.33	£30.00	W B
19	£3	1.05	20.00	£3.15	£60.00	W B

Figure 2: A Typical Lottery in the Risk Task

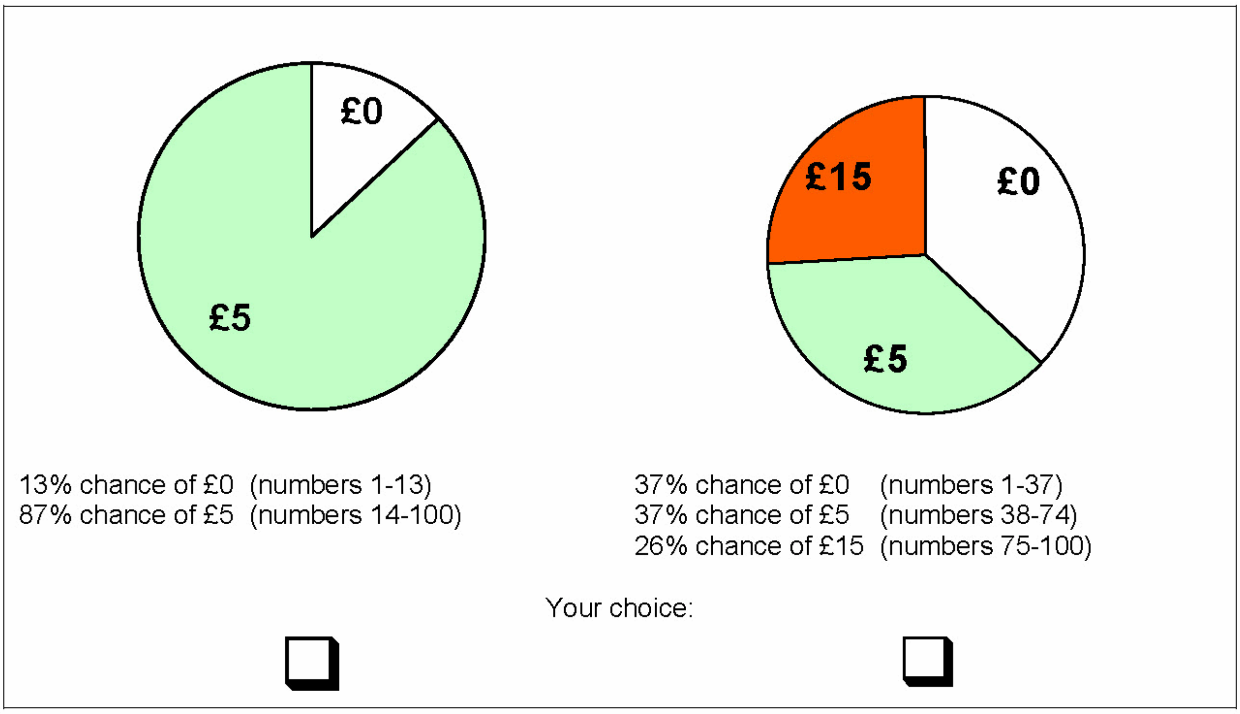


Figure 3: Comparison of our Findings with Griffin and Tversky (1992)

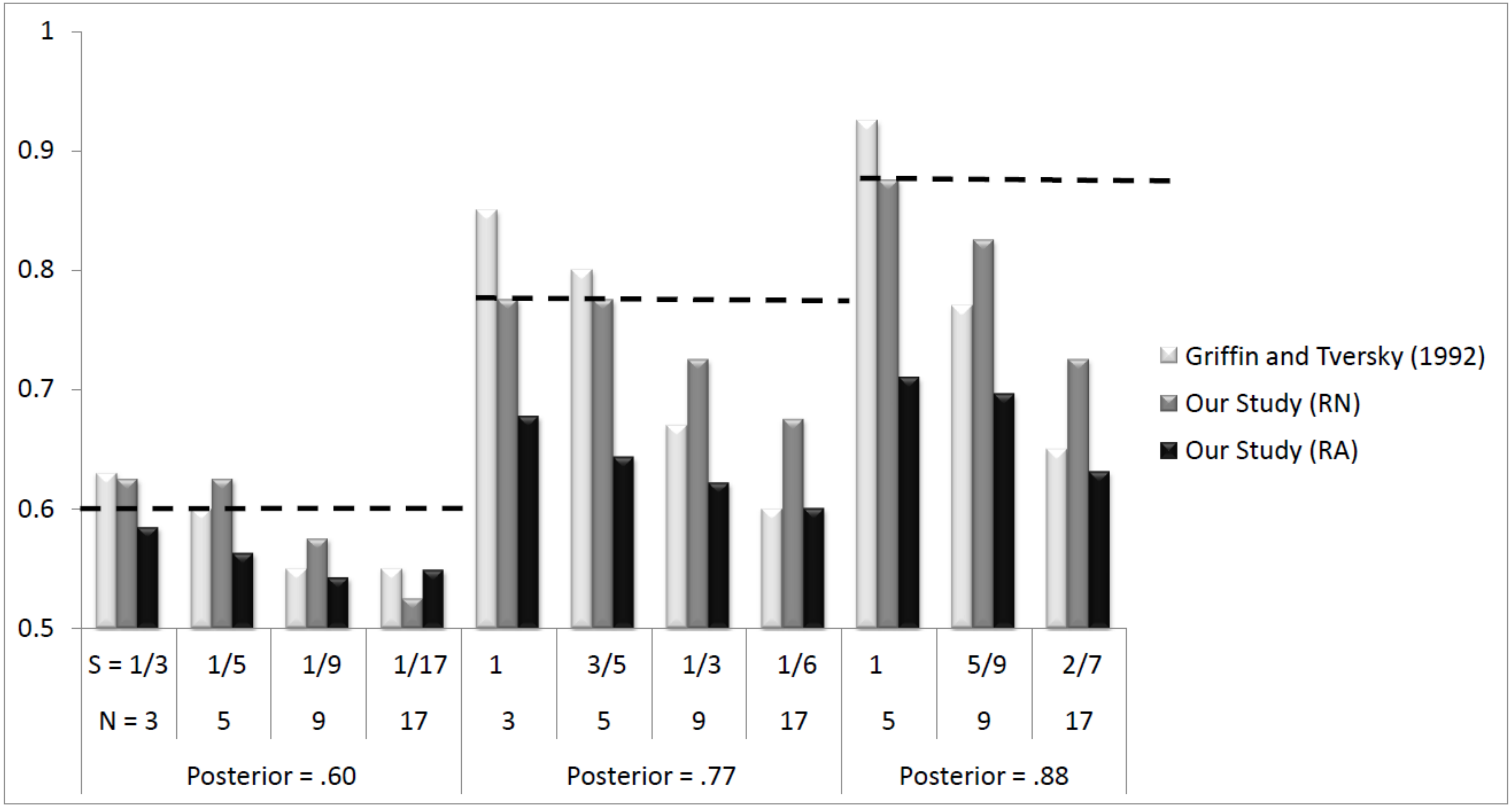


Figure 4: The Effect of Risk Aversion on Choice Probabilities

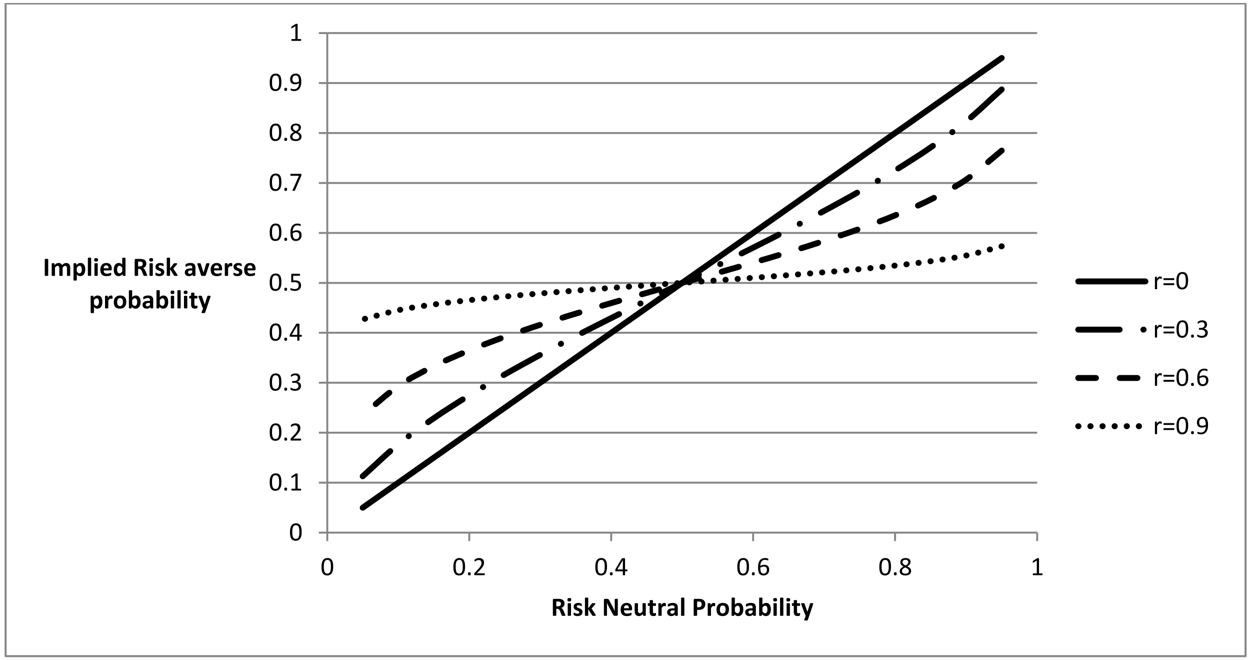


Table 1: Demographics (N = 111)

Age	
Mean	21.35
Median	20
Sex	
Male	58
Female	53
Field	
Economics, Finance, Business Administration	24
Engineering	4
Biological sciences, Health Medicine	6
Math, Computer or Physical Sciences	26
Social Sciences	23
Law	8
Psychology	4
Modern Languages	8
Other fields	8
Level of study	
Undergraduate	88
Postgraduate	12
Graduate	11
Mark at Bachelor degree	
Above 70% (first class)	30
between 60 and 69% (2.1)	72
between 50 and 59% (2.2)	5
No grades yet awarded	4

Table 2: Estimated Choice Probabilities

Bayesian Posterior	Sample(w, b)	Weight	Strength ratio	Choice-Probability (RN)	Std. Error (RN)	Choice-Probability (RA)	Std. Error (RA)	Frequency
Panel A: w>b								
0.88	(5,0)	5	1	0.824	0.027	0.700	0.025	45
0.88	(7,2)	9	0.56	0.808	0.023	0.688	0.020	53
0.88	(11,6)	17	0.29	0.732	0.023	0.631	0.017	81
0.77	(3,0)	3	1	0.793	0.020	0.678	0.017	55
0.77	(4,1)	5	0.6	0.731	0.016	0.634	0.634	227
0.77	(6,3)	9	0.33	0.694	0.027	0.602	0.022	95
0.77	(10,7)	17	0.18	0.682	0.017	0.61	0.014	73
0.6	(2,1)	3	0.33	0.643	0.013	0.583	0.010	154
0.6	(3,2)	5	0.2	0.601	0.009	0.553	0.007	394
0.6	(5,4)	9	0.11	0.566	0.016	0.533	0.010	87
0.6	(9,8)	17	0.06	0.568	0.020	0.55	0.015	44
Panel B: b>w								
0.88	(0,5)	5	1	0.861	0.017	0.715	0.024	72
0.88	(2,7)	9	0.56	0.821	0.012	0.699	0.014	110
0.88	(6,11)	17	0.29	0.726	0.024	0.63	0.021	47
0.77	(0,3)	3	1	0.785	0.017	0.676	0.016	77
0.77	(1,4)	5	0.6	0.763	0.011	0.652	0.011	218
0.77	(3,6)	9	0.33	0.734	0.013	0.638	0.012	112
0.77	(7,10)	17	0.18	0.664	0.013	0.593	0.010	106
0.6	(1,2)	3	0.33	0.655	0.014	0.586	0.012	110
0.6	(2,3)	5	0.2	0.631	0.008	0.572	0.007	412
0.6	(4,5)	9	0.11	0.591	0.013	0.549	0.009	117
0.6	(8,9)	17	0.06	0.579	0.014	0.548	0.010	82

Table 3: The Griffin and Tversky (1992) Model Assuming Risk Neutrality

In this table we present estimates for the coefficients on strength and weight, as per Equation (2), using a truncated regression model. We use this specification because the dependent variable is bounded between -1.399 and +2.201. In the truncated regression model we set the lower and upper limits equal to these values. Weight is the sample size, i.e., the number of dice rolled, and strength is calculated as abs(N_{white}-N_{blue})/$N_{total.}$

	Coefficient	Robust std. err.	P>|z|	95% Conf. interval	
constant	1.220	0.294	0.000	0.643	1.797
strength	1.921	0.525	0.000	0.891	2.951
weight	0.945	0.276	0.001	0.404	1.485
N	2786				
Tests	**Prob.>χ^2**				
constant=0	0.000				
strength=1	0.080				
weight=1	0.841				
strength=weight	0.001				

Table 4: Strength, Weight and the Variance of Subjective Beliefs

In this table we present estimates for the coefficients on strength and weight as per Equation (3) using a maximum likelihood hurdle model. We use this specification because the distribution of the dependent variable is highly non-normal. We estimate a probit model on 0 and 1, where 1 indicates that the dependent variable is greater than some threshold value (set equal to 0) and a constrained OLS on the dependent variable conditional on being greater than that threshold. We take the logarithmic transformation of the dependent variable when it assumes a value greater than the threshold. Weight is the sample size, i.e., the number of dice rolled, and strength is calculated as abs(N_{white}-N_{blue})/$N_{total.}$ In this test *n*=773.

	Coefficient	Robust std. err.	P>|z|	95% Conf. interval	
Panel A: Probit					
constant	0.866	0.149	0.000	0.575	1.158
strength	-0.727	0.183	0.000	-1.085	-0.369
weight	-0.042	0.011	0.000	-0.063	-0.020
Marginal effects					
constant	0.151	0.014	0.000	0.123	0.180
strength	-0.251	0.061	0.000	-0.370	-0.132
weight	-0.012	0.002	0.000	-0.016	-0.008
Panel B: Constrained OLS					
constant	-2.663	0.114	0.000	-2.887	-2.244
strength	-0.359	0.160	0.025	-0.673	-0.044
weight	-0.010	0.007	0.174	-0.025	0.004
Marginal effects					
constant	-0.065	0.007	0.000	-0.078	-0.051
strength	-0.021	0.010	0.032	-0.040	-0.002
weight	-0.001	0.0005	0.222	-0.002	0.0004
Joint significance	**Prob.>χ^2**				
strength	0.000				
weight	0.000				

Table 5: The Cost of the Strength/Weight Effect

Weight	Strength ratio	Probabilities: Bayes	Probabilities: OS	Probabilities: GT	Odds expressed as probabilities (Bookies): 0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	Cost: OS	Cost: GT
5	1	0.88	0.88													
				0.93									**GT**			1.96%
9	7/9	0.88	0.83									**OS**			2.41%	
				0.77							**GT**	**GT**				7.20%
17	12/17	0.88	0.73							**OS**	**OS**	**OS**			13.41%	
				0.65				**GT**	**GT**	**GT**	**GT**	**GT**				28.45%
3	1	0.77	0.78													
				0.85							**GT**	**GT**				6.01%
5	4/5	0.77	0.78							**OS**					1.07%	
				0.80												
9	6/9	0.77	0.73							**OS**					1.07%	
				0.67					**GT**	**GT**						4.27%
17	10/17	0.77	0.68						**OS**	**OS**					4.27%	
				0.60				**GT**	**GT**	**GT**						8.96%
3	2/3	0.60	0.63													
				0.63												
5	3/5	0.60	0.63													
				0.60												
9	5/9	0.60	0.58				**OS**								0.02%	
				0.55			**GT**									0.02%
17	9/17	0.60	0.53			**OS**	**OS**								1.99%	
				0.55			**GT**									0.02%

Table 6: SEU and the Griffin and Tversky (1992) Model

In this model we present estimates for the coefficients on strength and weight, as per Equation (4), using a structural maximum likelihood model that jointly estimates choice probabilities and risk attitudes, using data from both the Risk and the Belief task. r is the coefficient of risk aversion assuming a CRRA utility function of the form $U(x)= x^{1-r}/1-r$. *mu* is a structural noise parameter. Appendix 4 discusses in detail the estimation procedure. Weight is the sample size, i.e., the number of dice rolled, and strength is calculated as $abs(N_{white}-N_{blue})/N_{total.}$

	Coefficient	Robust std. err.	95% Conf. interval	
r	0.567	0.030	0.508	0.625
mu	0.208	0.029	0.151	0.264
strength	0.736	0.043	0.652	0.821
weight	0.442	0.040	0.363	0.520
Tests	Prob.>F			
strength=1	0.000			
weight=1	0.000			
strength=weight	0.000			

Appendix 1: Derivation of Equation 1

Here we provide the general formula for computing the posterior probability that a given set of dice (White or Blue, *W* or *B*) was chosen given the sample outcome *(w,b)*. The posterior probability, π, that *W* was chosen is:

$$\pi = p(W|w,b) = \frac{p(w,b|W)p(W)}{p(w,b)} \,. \tag{1}$$

The posterior probability of the Blue-set is then $1-\pi$. The likelihoods are the probabilities of given data, in this case (w,b), given the hypothesis, in this case W or B. The likelihood of W and B are therefore:

$$p(w,b|W) = [N!/(w!\, b!)]\, p(W)^{w}\, (1-p(W))^{(b)},$$

$$p(w,b|B) = [N!/(w!\, b!)]\, p(B)^{w}\, (1-p(B))^{(b)}, \tag{2}$$

The odds ratio is $p(w,b|W)/p(w,b|B)$ which with the help of the fact that $p(B) = 1-p(W)$, reduces with some simple algebra to the following:

$$L = \left(\frac{p(W)}{1-p(W)}\right)^{w-b} . \tag{3}$$

To separate out the effects of strength and weight we first take the log on both sides of the equation:

$$log(L) = (w-b)log\left(\frac{p(W)}{1-p(W)}\right) = N\left(\frac{w-b}{N}\right)log\left(\frac{p(W)}{1-p(W)}\right) \tag{4}$$

Taking the log on both sides of this equation gives the expression for weight and strength in Griffin and Tversky (1992):

$$log\left(log(L)/log\left(\frac{p(W)}{1-p(W)}\right)\right) = \log(N) + log\left(\frac{w-b}{N}\right) \tag{5}$$

Multiplying the two right-hand terms by α and β, respectively, we generate expression (1) in the text:

$$log\left(log(L)/log\left(\frac{p(W)}{1-p(W)}\right)\right) = alog(N) + \beta log\left(\frac{w-b}{N}\right) \tag{6}$$

Bayes Rule holds iff $\alpha=\beta=1$. If the number of *b* outcomes is higher than the number of *w* outcomes, then by the same process we obtain the following log odds ratio:

$$L = \left(\frac{p(W)}{1-p(W)}\right)^{b-w} , \tag{7}$$

and the following expression for weight and strength:

$$log\left(log(L)/log\left(\frac{p(W)}{1-p(W)}\right)\right) = alog(N) + \beta log\left(\frac{b-w}{N}\right) \tag{8}$$

Appendix 2: Instructions for the Belief Elicitation Task

In this stage of the experiment you will be betting on the outcomes of uncertain events. Usually we bet on events like football matches or elections, but in this task the events will be random choices made by the experimenter between two boxes, one blue and the other white. The experimenter will not tell you which box was chosen. At the start each box will have the same chance of being chosen, but once it has been chosen the experimenter will give you some information to help you work out the chances that it was blue or white. Armed with this information, you will make bets on which box was chosen.

The procedure, which is summarized on the accompanying picture, is as follows. The experimenter will first choose the box by rolling a 6-sided die with three blue and three white sides. If blue comes up he will choose the blue box, if white comes up he will choose the white one.

Both the white and blue boxes contain several dice, each having 10 sides. Both boxes have the same number of dice, which will vary over the course of the experiment. The dice in the blue box always have 6 blue sides and 4 white ones, while those in the white box have 4 blue sides and 6 white ones.

The experimenter will roll all the dice in the chosen box and tell you how many blue and white sides came up. He will not tell you which box was chosen.

Because the dice in the blue box have more blue sides than those in the white box, knowing the number of blue and white sides that come up can help you work out the chances that each box was chosen. For example, if more blue sides come up this means it is more likely to be the blue box, and if more white sides come up it is more likely to be the white box.

Once you have the information about the dice rolls, you will then make bets on which box was chosen.

About betting

You will be making bets with several betting houses or "bookies," just as you might bet on a football game or a horse race.

To familiarize you with betting, we will illustrate how it works with the example of a horse race.

Imagine a two horse race between Blue Bird and White Heat. Several bookies offer different odds for both horses. The table below shows the odds offered by three bookies along with the amounts they would pay if you staked £10 on the *winning* horse. The earnings are calculated by multiplying the odds by the stake. In this experiment you will be making bets on which box was chosen using a table like this. **At this point you should take some time to study the table.**

		Odds offered		**Earnings including the stake of £10**	
Bookie	**Stake**	**Blue Bird**	**White Heat**	**Blue Bird**	**White Heat**
A	£10	5.00	1.25	£50.00	£12.50
B	£10	3.33	1.43	£33.33	£14.30
C	£10	2.00	2.00	£20.00	£20.00

Below are three important points about betting.

1. **Your belief about the chances of each outcome is a personal judgment that depends on information you have about the different events.** For the horse race, you may have seen previous races or read articles about them. In the experiment the information you have about whether the blue or white box was chosen will be how many blue and white faces came up.
2. **Even if you believe Event X is more likely to occur than Event Y, you may want to bet on Y because you find the odds attractive.** For example, even if you believe White Heat is most likely to win you may want to bet on Blue Bird because you find the odds attractive. To illustrate, suppose you personally believe that Blue Bird has a 40% chance of winning and White Heat has a 60% chance of winning. This means that if you bet £10 on Blue Bird with Bookie A you believe there is a 40% chance of receiving £50.00 and a 60% chance of receiving nothing. You may find this more attractive than betting on White Heat, which you believe offers a 60% chance of 12.50 and a 40% chance of nothing.
3. **Your choices might also depend on your willingness to take risks or to gamble.** There is no right choice for everyone. In a horse race you might want to bet on the long-shot since it will bring you more money if it wins, but you also might want to bet on the favourite since it is more likely to win something.

For each bookie, whether you would choose to bet on Blue Bird or White Heat will depend on three things: your judgment about how likely it is each horse will win, the odds offered by the bookie, and how much you like to gamble or take risks.

Your choices

Now you are familiarized with odds, we can go back to the experimental betting task. Recall that the experimenter will first make a random choice of a blue or white box. Then he will roll the dice in the chosen box and tell you how many white and blue sides came up. Then you will consider the chances that the box chosen was blue or white, and make a series of bets.

You have a booklet of record sheets. Each record sheet shows the bookies you will be dealing with, and the odds they offer. There are 19 bookies on each sheet, and each offer different odds for the two outcomes. **Take a minute to look at one such record sheet, shown on the next page.**

There will be 30 separate events, and 19 bookies offer odds for each event. **You will make bets at all 19 bookies for all 30 events.**

For each bet, you have a £3 stake, and the record sheet shows the payoffs you will receive if you bet on the box that was actually chosen.

There is a separate record sheet for each of the 30 events. On each sheet you should circle W or B to indicate the bet you want to make with **all 19 bookies**.

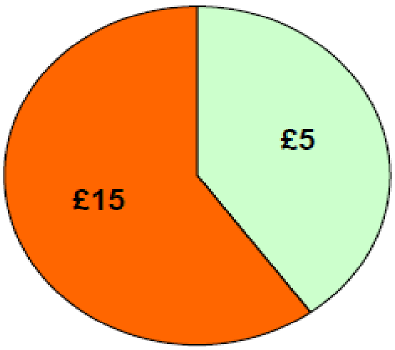
One and only one of the bets in the entire experiment will pay off for real. Therefore, please consider each bet as if it is the only one that will be paid out. After you have placed all your bets, you will roll a 30-sided die to determine which event will be played out, and a 20-sided die to determine which bookie will determine your earnings.

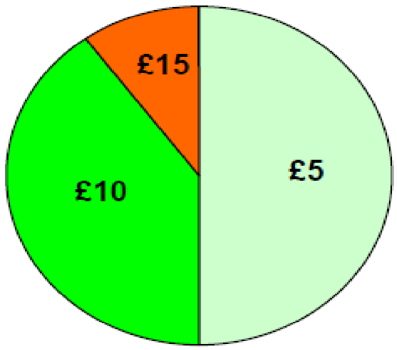
All payoffs are in cash, and are in addition to the £5 show-up fee that you receive just for being here.

Appendix 3: Instructions for the Risk Elicitation Task

This stage is about choosing between lotteries with varying prizes and chances of winning. You will be shown a series of 20 lottery pairs, and you will choose the lottery you prefer from each pair. You will actually get the chance to play one of the lotteries you choose, and will be paid according to the outcome of that lottery, so you should think carefully about your preferences.

Here is an example of one lottery pair. You will have to think about which lottery you would prefer to play and tick the appropriate box below





40% chance of £5	(numbers 1-40)	50% chance of £5	(numbers 1-50)
60% chance of £15	(numbers 41-100)	40% chance of £10	(numbers 51-90)
		10% chance of £15	(numbers 91-100)

Your choice:

☐ ☐

The outcome of the lotteries will be determined by the draw of a random number between 1 and 100. We will ask you to roll a 100-sided die that is numbered from 1 to 100, and the number on the die will determine the outcome of the lotteries.

In the above example the left lottery pays five pounds (£5) if the number on the die is between 1 and 40, and it pays fifteen pounds (£15) if the number is between 41 and 100. The light green segment of the pie chart corresponds to 40%, and the orange segment corresponds to 60% of the area.

Now look at the pie chart on the right. It pays five pounds (£5) if the number drawn is between 1 and 50, ten pounds (£10) if the number is between 51 and 90, and fifteen pounds (£15) if the number is between 91 and 100. As with the lottery on the left, the pie slices represent the fraction of the possible numbers which yield each payoff. For example, the size of the £15 pie slice is 10% of the total pie.

Each of the 20 lottery pairs will be shown on a separate sheet of paper. On each sheet you should indicate your preferred lottery by ticking the appropriate box. After you have worked through all the lottery pairs, please raise your hand. You will then roll a 20-sided die to determine which pair of lotteries will be played out, and then roll the 100-sided die to determine the outcome of the chosen lottery.

For instance, suppose you picked the lottery on the left in the above example. If you roll the 100-sided die and the number 37 is shown, you would win £5; if it was 93, you would get £15. If you picked the lottery on the right and drew the number 37, you would get £5; if it was 93, you would get £15.

Therefore, your payoff is determined by three things:

- which lottery pair is chosen to be played out using the 20-sided die;
- which lottery you selected, the left or the right, for the chosen lottery pair; and
- the outcome of that lottery when you roll the 100-sided die.

This is not a test of whether you can pick the best lottery in each pair, because none of the lotteries are necessarily better than the others. Which lotteries you prefer is a matter of personal taste.

Please work silently, and think carefully about each choice.

All payoffs are in cash, and are in addition to the £5 show-up fee that you receive just for being here.

Appendix 4: Joint Estimation of Risk Aversion and Choice Probabilities

We start by explaining the econometric analysis of the data collected in the risk task. We assume a Constant Relative Risk Aversion (CRRA) utility function, defined as:

$$U(y) = y^{(1-r)}/1-r, \tag{1}$$

where *r* is a parameter to be estimated, and *y* is income from the experimental choice. The utility function (1) can be estimated using the responses from our risk task using maximum likelihood and a latent EUT structural model of choice. In the lotteries provided there are *K* possible outcomes, where $K \leq 3$. The objective probabilities for each outcome, p_k, are illustrated in both pie chart and textual form, as shown by Figure 3. Therefore, Expected Utility is simply the probability weighted utility of each outcome in each lottery *i*:

$$EU_i = \Sigma_{k=1,K} [p_k \times u_k]. \tag{2}$$

The EU for each lottery pair (the left and the right, as shown by Figure 3) is calculated for a candidate estimate of *r*, defining the index:

$$\nabla EU = EU_R - EU_L \tag{3}$$

where EU_L is the expected utility of the "left" lottery and EU_R is the expected utility of the "right" lottery. This latent index, based on latent preferences, is then linked to the observed choices using a standard cumulative normal distribution function $\Phi(\nabla EU)$. This "probit" function takes any argument between $\pm\infty$ and transforms it into a number between 0 and 1.

Thus we have the probit link function:

$$\text{prob(choose lottery R)} = \Phi(\nabla EU) \tag{4}$$

An important extension of the core model is to allow for respondents to make some errors. We use the contextual error specification proposed by Wilcox (2010). It posits the latent index:

$$\text{prob(choose lottery R)} = \Phi [(\nabla EU)/v)/\mu] \tag{5}$$

instead of (5), where *v* is a normalizing term for each lottery pair L and R, and $\mu>0$ is a structural "noise parameter" used to allow some errors from the perspective of the deterministic EUT model. This parameter is defined as the maximum utility over all prizes in this lottery pair minus the minimum utility over all prizes in *this* lottery pair, and ensures that the *normalized* EU difference $[(EU_R - EU_L)/v]$ remains in the unit interval. As $\mu \to \infty$ this specification collapses ∇EU to 0 for any values of EU_R and U_L, so the probability of either choice converges to ½. Therefore, a larger μ means that the difference in the EU of the two lotteries, conditional on the estimate of *r*, is less predictive of choices. Additional details of the estimation methods used, including corrections for "clustered" errors when we pool choices over respondents and tasks, are provided by Harrison and Rutström (2008; p.69).

Thus the likelihood of the observed responses, conditional on the assumptions of EUT and CRRA, depends on the estimates of *r* given the above statistical specification and the observed choices. The log-likelihood is then

$$\ln L(r, \mu; y, \mathbf{X}) = \Sigma_i [(\ln \Phi(\nabla EU)\times\mathbf{I}(y_i = 1)) + (\ln (1-\Phi(\nabla EU))\times\mathbf{I}(y_i = -1))] \tag{6}$$

where $\mathbf{I}(\cdot)$ is the indicator function, y_i =1(-1) denotes the choice of the Option R (L) lottery in risk aversion task *i*, and **X** is a vector which contains ν and any other data, such as demographics and task characteristics.

We now explain how we use the choices in the belief task to construct the second likelihood to estimate choice probabilities. As shown in Figure 2, the subject that selects event W from a given betting house receives

$$EU_W = \pi_W \times U(\text{payout if W} \mid \text{bet on W}) + (1-\pi_W) \times U(\text{payout if B} \mid \text{bet on W}) \quad (7)$$

where π_w is the subjective probability that W will occur. The payouts that enter the utility function are defined by the odds that each bookie offers. We similarly define the EU received from a bet on event B as the complement of event A:

$$EU_B = \pi_W \times U(\text{payout if W occurs} \mid \text{bet on B}) + (1-\pi_W) \times U(\text{payout if B occurs} \mid \text{bet on B}). \quad (8)$$

We observe the bet made by the subject for a range of odds, so we can calculate the likelihood of that choice given values of r, π_w and μ, again assuming *EUT* and CRRA.
The rest of the structural specification is exactly the same as for the choices over lotteries with objective probabilities. Thus the likelihood function for the observed choices in the belief task is:

$$\ln L(r, \pi_W, \mu; y, \mathbf{X}) = \Sigma_i \left[(\ln \Phi(\nabla EU) \times \mathbf{I}(y_i = 1)) + (\ln (1-\Phi(\nabla EU)) \times \mathbf{I}(y_i = -1)) \right] \quad (9)$$

The *joint estimation* problem is to find values for r, π_w and μ that maximize the *sum* of (6) and (9).

Using this joint estimation framework we can estimate the model in (1), controlling for risk attitude. We can replace π_w with two parameters, α and β, that define π_w in terms of the strength and weight of the sample data, as shown by (1). For the case in which w>b we specify:

$$\gamma = \alpha \log N + \beta \log((w-b)/N) \quad (10)$$

$$\lambda = \exp[\exp(\gamma) \exp(0.6/0.4)] \quad (11)$$

By re-arrangement we obtain:

$$\pi = 1/(1+(1/\lambda)) \quad (12)$$

A similar specification when b<w yields:

$$\pi = 1/(1+\lambda) \quad (13)$$

Using these derivations we can express π_w in the maximum likelihood model in terms of α and β, and then estimate these parameters using the joint estimation procedure explained above.

Endnotes

[1] As discussed later, to facilitate hypothesis testing Griffin and Tversky (1992) used a transformation of this ratio *2h/N-1* when conducting regression analyses.
[2] These doubts are echoed by Gilboa, Postlewaite and Schmeidler (2008): Suppose that Ann would like to elicit her subjective probability for the event "Candidate X will graduate successfully from the program." She might ask herself questions: *2h/N-1* when conducting regression analyses.
[2] These doubts are echoed by Gilboa, Postlewaite and Schmeidler (2008): Suppose that Ann would like to elicit her subjective probability "Do I prefer to bet on X graduating vs. failing at odds 1:2?" *Such preference questions may have a more palpable meaning to Ann than the question, "What is the precise probability p that X will graduate?*" (p. 178, italics added)
3 Evidence of the difference between stated-probabilities and choice-probabilities can be found in Costa-Gomes and Weizsäsker (2008), and Rutström and Wilcox (2009).
4 Berman (2007) asked subjects how they answered a series of such questions, and found that over 30% explicitly reported they were using this "ratio heuristic". Less than 4% described anything that could be interpreted as the use of Bayes rule.
[5] The most unlikely pattern is (5,0) with probability of occurring equal to 0.044, followed by the pattern (7,2) with probability 0.091, and the pattern (11,6) with probability 0.104.
[6] Hence a subject whose true switch point would be with bookies 1 through 9 would switch at a higher-numbered bookie, no higher than bookie #10, and a subject whose true switch point would be with bookies 10 to 19 would switch at lower-numbered bookies no lower than bookie #10.
[7] Table 2 does not show 25 dice combinations that emerged and that did not have equivalents in the original Griffin and Tversky (1992) design. These include six instances of (13W,4B), four of (12W,5B), one of (14W,3B), one of (9W,0B), one of (8W,1B), four of (5W,12B), four of (4W,13B), one of (3W,14B), two of (1W,8B) and one pattern of (2B,15W).
[8] The dependent variable in the regression is bounded between -1.399 and +2.201. These are the values obtained when elicited probabilities equal 97.5 and 2.5, respectively, which are the minimum and maximum values allowed in our design. In the truncated regression model we set the lower and upper limits equal to these values.
[9] In total we have 2,970 estimates of subjective probabilities (including the samples that were omitted from Table 2 as per endnote 5). We lose 184 observations when we make the necessary transformations to estimate the model in Equation 2. These observations correspond to subjects who, despite observing more blue dice in the sample, placed bets that imply the white cup was more likely. Our results are almost identical when we exclude the samples not included in Griffin and Tversky's (1992) design.
[10] See also relevant discussions regarding the reductions of "behavioural biases" in experimental economics settings in Grether and Plott (1979), Grether (1980) and Plott and Zieler (2005).
[11] This procedure is criticized by Engelberg, Manski and Williams (2009), who argue instead for the elicitation of subjective densities themselves. The point was also made forcefully by Zarnowitz and Lambros (1987) with respect to the uncertainty of a sample of point forecasts.
[12] Specifically we estimate a probit model on 0 and 1, where 1 indicates that the dependent variable is greater than some threshold value and a constrained OLS on the dependent variable conditional on being greater than that threshold. The procedure yields estimates for the coefficients of the dependent variables for each of the two models. With a joint maximum likelihood implementation of both estimation stages, any estimates are also efficient. These estimates can be used to calculate total and marginal effects. We take the logarithmic transformation of the dependent variable when it assumes a value greater than the threshold. We set the threshold equal to 0, which results to 60% of the observations being above the threshold.
[13] For more evidence on the effect of non-risk neutral preferences on subjective beliefs see Antoniou, Harrison, Lau and Read (2012), which uses the same raw data as our study but is concerned with the structural estimation of beliefs and preferences in a Bayesian context, considering both EUT and non-EUT preferences.