A new and much improved version of the first experiment reported in this paper is published in:

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

Proper scoring rules provide convenient and highly efficient tools for incentive compatible elicitations of subjective beliefs. As traditionally used, however, they are valid only under expected value maximization. This paper shows how they can be generalized to modern ("nonexpected utility") theories of risk and ambiguity, yielding mutual benefits: people using proper scoring rules can benefit from the empirical realism of nonexpected utility, and people analyzing ambiguity attitudes can benefit from the efficient measurements through proper scoring rules. An experiment demonstrates the feasibility of our generalized proper scoring rule.

We do not intend to publish this working paper anymore, but it will stay accessible on internet (because it has already attracted some citations).

If you are primarily interested in the first experiment, please cite the article mentioned above.

Of course, you are welcome to cite this working paper for the second experiment.

Joep Sonnemans

(http://www1.fee.uva.nl/creed/people/sonnemans/index.shtml)

Theo Offerman

(http://www1.fee.uva.nl/creed/people/offerman/index.shtml)

Is the Quadratic Scoring Rule really incentive compatible?

Joep Sonnemans

Theo Offerman

December 2001

Abstract
Theoretically, the quadratic scoring rule (QSR) is an incentive compatible mechanism to elicit beliefs. When rewarded by a QSR, respondents should report their true subjective beliefs if they want to maximize expected value. But there is a possibility that respondents misreport their true beliefs as a consequence of their risk attitude or as a consequence of probability weighting. The use of the QSR is nowadays widespread in experimental economics, but behavioral properties of this mechanism have received only little attention so far. We present a method that corrects subjects' reported beliefs for undesired effects of risk attitudes and probability weighting. We report the results of two methodological experiments on the QSR. In the first experiment, we examine whether the incentives provided by the QSR encourage subjects to report their true subjective beliefs. In the second experiment, we investigate the effects of salient incentives on the effort that subjects tend to exert and on their performance when they formulate their beliefs. We find that subjects tend to report their true subjective beliefs when rewarded by a QSR. However, subjects also exert substantial effort and make good judgments when they are rewarded with a flat fee.

Key-words: Quadratic scoring rule, experiment.

JEL C90, D84.

Address:
University of Amsterdam
Faculty of Economics and Econometrics
CREED
Roetersstraat 11
1018 WB Amsterdam
the Netherlands
e-mail: joeps@fee.uva.nl, theoo@fee.uva.nl

* We are grateful for comments by Tanga McDaniel, Peter Wakker and participants of the European meeting of the Economic Science Association in Amsterdam, October 2000. Theo Offerman is grateful for financial support from the Netherlands' Organization for Scientific Research.

1. Introduction

In many situations, a researcher would like to have precise information about beliefs of economic agents. One interesting possibility is to ask agents to report their beliefs. When doing so, a researcher has to make a decision whether to reward agents for stating their beliefs or not. Without salient incentives, there is a danger that agents do not take their task seriously and that their reports are noisy. A researcher may therefore opt to use one of the incentive mechanisms discussed in the literature (Murphy and Winkler, 1970; Savage, 1971; Holt, 1986). Agents who are rewarded with a payoff generated by a strictly proper scoring rule will truthfully reveal their beliefs when they maximize expected value (or expected utility with a linear utility function).

Murphy and Winkler (1970) discuss three strictly proper scoring rules: the logarithmic, the spherical and the quadratic scoring rule. Of these, the Quadratic Scoring Rule (QSR) has been used most frequently. Suppose that the agent reports the discrete probability distribution $p=(p_1,..,p_n)$ where p_i $(1\leq i\leq n)$ represents the reported probability that event i occurs. If the event j occurs, the QSR offers the agent a payoff $Q_j(p)$ equal to:

$$Q_j(p) = a + 2b * p_j - b * \sum_{i=1}^{n} (p_i)^2 \qquad (1)$$

The parameters a and b are chosen by the researcher. The minimum payoff is a-b and the maximum payoff is a+b. If a is set equal to b, the minimum payoff is 0 and the maximum payoff is 2a. Note that the QSR can even be used when the underlying real distribution is unknown to the researcher, e.g. a weather forecaster's estimate of the probability of rain on a certain day (see Camerer, 1995, for references), or the number of contributions in a public good game.

The QSR has been used by, for example, McKelvey and Page (1990) in an experiment on information aggregation, by Friedman and Massaro (1998) in an individual learning experiment and by Nyarko and Schotter (2000) in a study on belief learning in a strategic game with a unique mixed strategies equilibrium. Kraemer and Weber (2001) use the QSR in an experiment in which subjects sequentially process information about their predecessor's beliefs and, if they are willing to pay, some private information. McDaniel and Ruthström (2001) make use of the QSR in an individual problem solving experiment.[1] Huck and Weizsäcker (2002) elicit beliefs of

[1] This paper uses a slightly more complicated reward scheme. Subjects have to report probabilities for

one group of subjects for another group of subjects' choices between lotteries. They compare beliefs elicited via a QSR procedure with beliefs elicited via a Becker-DeGroot-Marshak pricing rule, and conclude that the QSR procedure yields more accurate beliefs. We have also used the QSR (Offerman, Sonnemans and Schram, 1996, 2001, Sonnemans, Schram and Offerman, 1998, 1999 and Offerman and Sonnemans 1998, Offerman, forthcoming) to elicit subjects' beliefs in public good games, individual decision tasks and the hot response game.

It is well known that scoring rules are not incentive compatible if subjects are not risk neutral or if they fall prey to probability weighting.[2] McKelvey and Page (1990) use a procedure that in theory induces subjects to behave as if they are risk neutral. They modify the payoff rule by offering lottery tickets instead of money. Each lottery ticket earned by a subject increases the probability of winning the high payoff instead of the low payoff. Because a utility function is linear in probabilities, this procedure should induce risk neutrality. However, it is doubted whether this procedure is successful in doing so (e.g., Davis and Holt, 1993, p.474). A recent study by Selten, Sadrieh and Abbink (1999) suggests that the procedure does not reach its goal. We present another method that allows the researcher to reconstruct the subjects' true subjective beliefs from their reported beliefs. This simple method corrects for undesired effects of both an agent's risk attitude and probability weighting. It is not restricted to the present specific experiment, but can be used in any setting where a researcher wants to elicit true subjective probabilities.[3]

In this paper, we experimentally investigate two methodological issues related to the use of scoring rules. Previous studies that made use of scoring rules did not correct reported beliefs for risk attitudes and probability weighting. We first address the question whether this was an important omission in those studies. In other words, do risk attitudes and probability weighting pose important empirical problems for the elicitation of beliefs with scoring rules? There is a theoretical reason why risk attitudes might not have such a pronounced effect in studies with repeated tasks. If agents repeatedly report beliefs and if they are only concerned about their final wealth, the law

---

10 intervals and get quadratic scoring rule payments for each interval (the QSR is used 10 times with n=2, instead of once with n=10).

[2] Probability weighting was first described in Kahneman and Tversky (1979). It refers to the tendency of subjects to overweight small probabilities and underweight large probabilities in the decision making process.

[3] This method was suggested to us by an anonymous referee of our paper on overreaction (Offerman

of large numbers ensures that they should act as if risk neutral. Thus, theoretically, problems caused by risk attitudes may be mitigated in repeated tasks. However, experimental evidence suggests that this large number effect may not hold in practice (see, for instance, Barron and Erev, 2000). The results obtained in our first experiment nevertheless suggest that subjects' reported beliefs are not biased by their risk attitudes or probability weighting when they are rewarded with the help of a QSR.

A second issue is whether subjects report their beliefs more seriously when they have incentives. Does the QSR motivate subjects to exert more effort on the task, and to formulate better beliefs? Some early psychological studies (Beach and Philips 1967, Jensen and Peterson 1973) suggest that there is no difference in performance between subjects motivated by a proper scoring rule and subjects who are not motivated by monetary incentives. A recent study by Friedman and Massaro (1998) on beliefs in a probability matching task does not find significant differences between paid and unpaid subjects. However, the latter paper suggests that relatively more of the unpaid subjects were unmotivated (in the sense that they repeatedly reported the default probability value 0). All these studies lack an independent measurement of the effort. In our second experiment, subjects estimate the proportion blue of a wheel of fortune. To improve their beliefs, subjects can ask up to 20 (time consuming) draws per wheel. In the flat fee treatment, subjects earn the same amount irrespective of their estimates, while subjects in the QSR treatment are paid according to a QSR. We investigate the hypotheses that in both treatments subjects will exert an equal amount of effort and will give equally good estimates. Both hypotheses cannot be rejected.

The organization of the remainder of the paper is as follows. Section 2 describes the procedures of experiment 1 and reports its results. Section 3 deals with experiment 2 and section 4 concludes.

2. Experiment 1: the effects of risk attitude and probability weighting

Design Experiment 1

Experiment 1 is designed to investigate whether risk attitudes and probability weighting bias subjects' reported beliefs. It consists of two parts. The first part

---

and Sonnemans, 2000).

replicates the experiment on overreaction reported in Offerman and Sonnemans (2000). We summarize the main features of that study. It examines the phenomenon of overreaction that is observed in both sports markets and financial markets. In both types of markets it is found that past losers outperform past winners. This effect has been attributed to recency (traders overweight recent information and underweight the base-rate) or to the hot hand effect (traders overestimate the auto-correlation in a series of draws: they more easily discover trends - hot hands - than a Bayesian observer would). The experiment is designed to distinguish between the two explanations.

The experimental setup is as follows. A coin is selected randomly from an urn containing an equal number of false and fair coins (the prior probability of a false coin = 0.5). A fair coin has no memory: each toss of the coin will be head (tail) with probability 0.5 (0.5). A false coin has the property that the previous outcome is repeated with probability 0.7. If the previous outcome was head, the new outcome will be head with probability 0.7 and it will be tail with probability 0.3. Thus, the outcome of the toss of a false coin depends only on the outcome of the previous toss. The outcome of the first toss with a false coin is head (tail) with probability 0.5 (0.5). The decision-maker is not told whether the randomly selected coin is fair or false. The coin is tossed twenty times yielding a series of heads and tails. The decision-maker observes the series and predicts the probability that the series was generated by a false coin.

The recency hypothesis predicts that subjects will overweight the information contained in the series of coin tosses and underweight the base rate information. The effect of recency depends on the number of alternations in the series of heads and tails generated by the coin. If this number is such that a Bayesian observer would report a higher probability than 50%, then neglect of the prior distribution would induce a decision-maker to overestimate the probability that the coin is false. On the other hand, if the number of alternations leads a Bayesian observer to predict a probability smaller than 50%, then neglect of the prior distribution would induce a decision-maker to underestimate the probability that the coin is false.

The hot hand hypothesis predicts that subjects overestimate the autocorrelation in the series. A series actually generated by a fair coin without autocorrelation will then be perceived as a series of a false coin with positive autocorrelation. A series actually generated by a false coin with positive autocorrelation will then be perceived

as a series of a false coin with even higher autocorrelation. In both cases, the decision-maker reports a higher than Bayesian posterior probability that the coin is false.

For each estimate that a coin is false a subject receives a payoff determined by a quadratic scoring rule with $a=b=5000$ points. Let R denote the reported probability of a false coin in percentages, then the payoff is $10000\text{-}R^2$ points if the coin is fair and $200*R\text{-}R^2$ points if the coin is false. At the end of the experiment the points are exchanged for money. Subjects do not know the formula of the scoring rule, but receive a payoff table on a hand out. The table displays the payoff for each (integer) estimate between 0% and 100% when the coin is fair and when the coin is false. The table is included in the appendix. The instructions explain that it is in the best interest of subjects to report their true beliefs (it is not our goal to test whether subjects will discover the properties of the mechanism by themselves.) The instructions contain some questions to check understanding.

In the second part, subjects make choices that help us to derive true subjective equivalents from reported probabilities. In total, subjects make 9 choices, each between 21 gambles. Each of the choices uses exactly the same 21 pairs of payoffs as those reported in table 1. The payoffs in the gambles are determined with the help of the quadratic scoring rule used in part 1. For each of the 21 gambles the first payoff is equal to $200*i\text{-}i^2$ and the second payoff is equal to $10000\text{-}i^2$, with i increasing from 0 to 100 in steps of 5. We present the gambles in the format shown in table 1. The 9 choices differ in the probabilities of the outcomes in the second and third column. Each element from the set {(10%,90%), (20%,80%), ..., (80%,20%), (90%,10%)} is used once, where the first (second) number of an element represents the probability of the outcome in the second (third) column. Note that the choice (80%,20%) is mathematically equivalent to the choice (20%,80%): only the framing differs, the two columns are swapped and the gambles are renamed. In principle, we have two observations for each choice with probability 60%, 70%, 80% and 90% (or similarly, for each choice with probability 40%, 30%, 20% and 10%). Therefore, we can compute the true subjective probability for a reported probability twice. This is a useful procedure if one assumes that there is some noise in the decision making process. If the two true subjective probabilities for some reported probability differ, we simply take the average as the best approximation of the true subjective probability.[4]

[4] Note that the risk profile is necessarily symmetric around the (50%,50%) point, as a result of the 'double counting' of the choices. If we do not double count the choices, results are very similar. The main

Here follows an example to illustrate how a subject's choice between gambles can help to map reported probabilities into true subjective probabilities. To derive what a subject would report for a true subjective probability of 80%, we offer the subject a choice from a set of 21 gambles (table 1). Each of these gambles offers the same probability of 80% on a first payoff (column 2 in the table) and 20% on a second payoff (column 3 in the table). A risk neutral subject should choose alternative E for this particular choice. Gambles A, B, C and D are the risky gambles and gambles F, G, ..., K are safe gambles for this particular choice. Assume that a risk averse subject chooses gamble G corresponding to 70%. If this subject reports a probability of a false coin of 70% in part 1, then we conclude that this report corresponds to a true subjective probability of 80%. Notice that it is not important whether a subject biases the reported probability distribution for risk or for probability weighting reasons. The procedure corrects in both cases.

In this way a risk profile can be established for each subject. We further explain this procedure with the help of the gambles selected by one of the subjects. Figure 1 shows the 9 choices of this subject. We add the points (0%,0%) and (100%,100%) to the gambles selected. We connect the 9+2 points in the graph linearly. Each reported probability in the first part of the experiment (displayed on the y-axis) is mapped into a corresponding true subjective probability (displayed on the x-axis) via this risk profile. For example, a reported probability of 70% would have been mapped into a 60%+10%*(70-62.5)/(77.5-62.5) = 65% true subjective probability for this subject.

The mapping procedure yields unique true subjective probabilities as long as the risk profile is strictly increasing. If the risk profile is horizontal at some part, we set the true subjective probability equal to the average probability of this horizontal part. This particular risk profile has horizontal parts on [0,10%] and on [90%,100%]. Thus, a reported probability of 100% would have been mapped into a true subjective probability of 95%. If the procedure described above leads to a risk profile that decreases on some intervals it does not make sense to map reported probabilities into true subjective probabilities for this subject, and we exclude the subject from the sample for non-monotonicity reasons.

---

difference is that in that case one additional subject has to be excluded for non-monotonicity reasons.

Table 1: choice between gambles in second part experiment 1 (this is one of the nine decision sheets)

Alternative:	If the outcome of your toss with a ten-sided die is 1 or 2 (20% probability), your payoff is:	If the outcome of your toss with a ten-sided die is 3, 4, 5, 6, 7, 8, 9 or 0 (80% probability), your payoff is:
A (100%)	0	10000
B (95%)	975	9975
C (90%)	1900	9900
D (85%)	2775	9775
E (80%)	3600	9600
F (75%)	4375	9375
G (70%)	5100	9100
H (65%)	5775	8775
I (60%)	6400	8400
J (55%)	6975	7975
K (50%)	7500	7500
L (45%)	7975	6975
M (40%)	8400	6400
N (35%)	8775	5775
O (30%)	9100	5100
P (25%)	9375	4375
Q (20%)	9600	3600
R (15%)	9775	2775
S (10%)	9900	1900
T (5%)	9975	975
U (0%)	10000	0

Notes: After the letter (representing the gamble) a percentage is displayed. A risk neutral person would select a specific gamble if the probability of the outcome of the last column equals the corresponding percentage of that gamble. In this specific decision sheet, a risk neutral person chooses E.The percentages after the letters were not shown in the experiment.

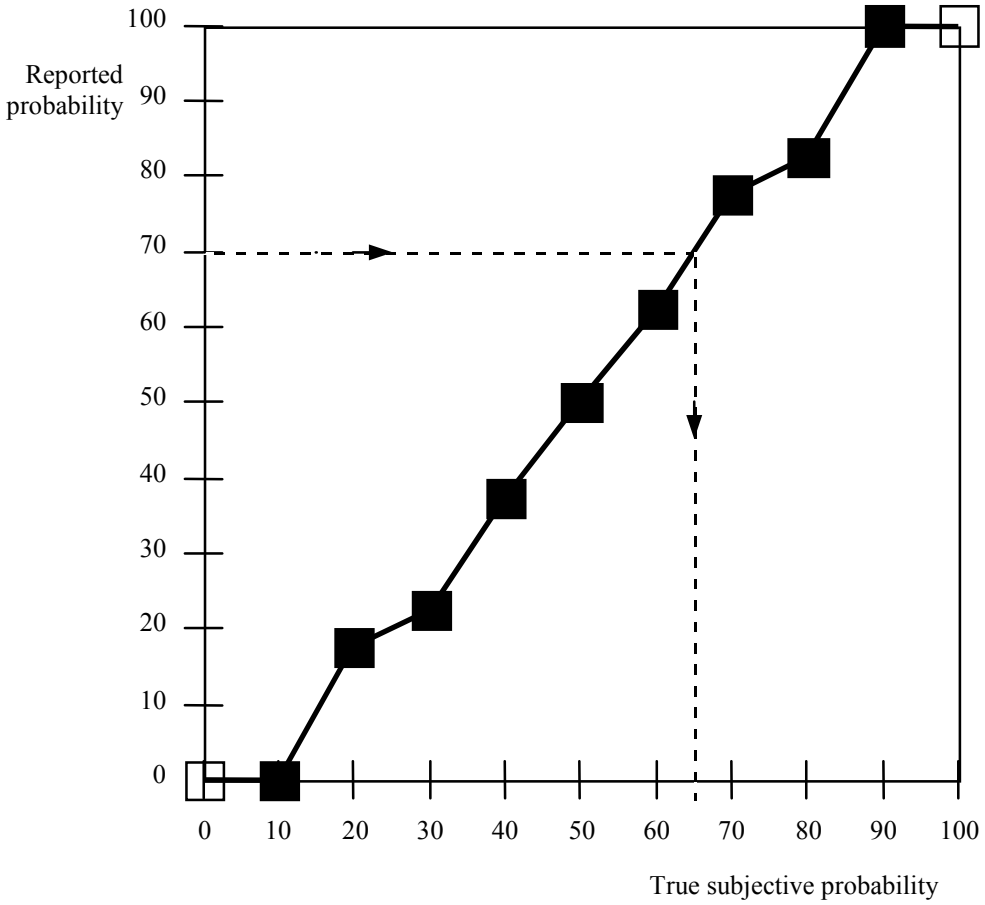

Figure 1: a subject's reported probabilities as function of the true subjective probabilities.

The first part of the experiment is computerized, but the second part is not. Subjects make the 9 choices of the second part by selecting their preferred gambles on paper. The choices are presented in a random order. The payoffs are again denoted in points. The same exchange rate is used as the one for the first part of the experiment (8000 points=1 guilder). When subjects have made all their decisions, they toss a ten-sided die 9 times, once for each choice. The outcome of the toss and the gamble selected by the subject determines her payoff for a particular choice. The nine payoffs are added to the amount earned in the first part of the experiment to determine a subject's total payoff in the experiment.

Results Experiment 1

In total 21 subjects participated: 14 of these subjects major in economics and 5 in other fields (2 did not report their study); 8 of these subjects are female and 13 are male. An average of 19.00 guilders was earned by subjects in the first part of the

experiment and 8.95 guilders in the second part of experiment 2 in 1 hour (1 guilder can be exchanged for approximately 0.45 US dollar). We exclude 3 subjects from the analyses for non-monotonicity reasons described above.

The results for the first part are in line with the results for the original experiment on overreaction. Subjects estimate the probability that a coin is false higher than a Bayesian observer would (on average 60.8%, 59.6% in the original overreaction experiment and 45.9% for the Bayesian observer). As expected by the hot hand hypothesis, subjects overestimate the probability of a false coin for both coins that seem fair and coins that seem false to a Bayesian observer.

Only 2 subjects exactly maximize expected value in the second part of the experiment. To investigate the possible biasing effects of risk attitude and probability weighting, we map reported probabilities into true subjective probabilities with the help of subjects' choices in the second part of the experiment as described in the previous section. In 79% of the decisions the true subjective probabilities differ 5 percent points or less from the reported probabilities and in 33% of the decisions the true subjective and reported probabilities match exactly. Figure 2 displays both the mean reported and the mean corresponding true subjective probabilities as function of the Bayesian probabilities. The two lines are almost indistinguishable. At the individual level we also observe only small differences. The mean absolute difference between true subjective and reported probabilities is for most subjects smaller than a few percentage points (the mean is 3.6 percentage points. The exception is one subject with a mean difference of about 10 percentage points). Deviations from truthtelling are not systematic. The scoring rule does not bias the results. Both reported and true subjective probabilities are best explained by a combined process of the hot hand effect and random noise. The average true subjective probability of a false coin is 60.3%, almost indistinguishable from the average reported probability (60.8%).

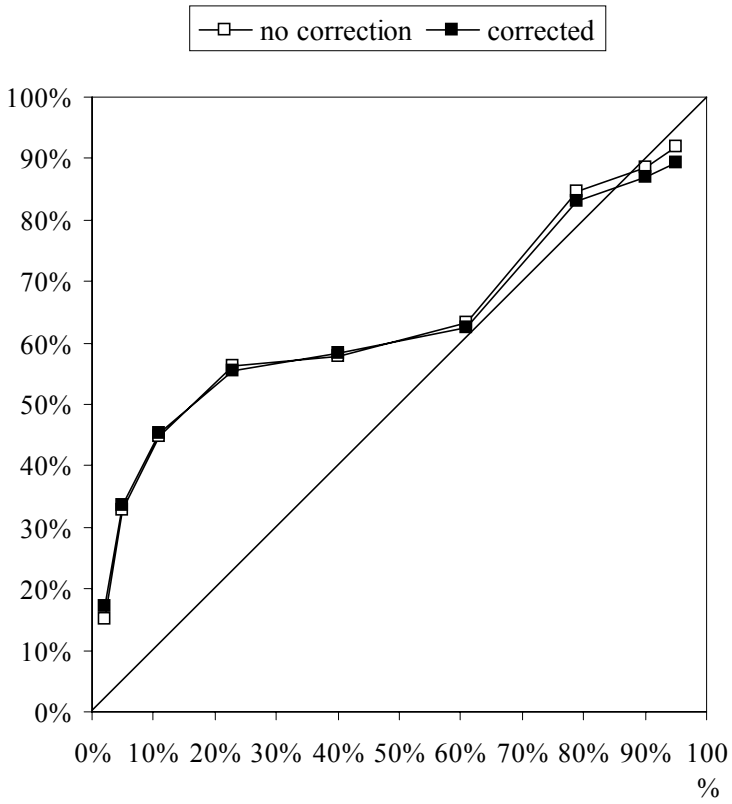

Figure 2: mean reported and mean true subjective probabilities as function of the Bayesian probabilities (Bayesian probabilities on the horizontal axis). The straight line R=B is added as a benchmark.

3. Experiment 2: motivational aspects

Experiment 2 focuses on the motivational aspects of the QSR. We try to answer two questions: when paid according to the QSR, do subjects (1) exert more effort on the task, and (2) make better judgments, compared with subjects who receive a flat fee.

Design Experiment 2

The second experiment consists of 20 periods. In each period subjects are confronted with a (new) wheel of fortune which is partly blue and partly yellow. The percentage blue is unknown to the subjects, but they know that all probabilities from 0% to 100% are equally likely. Subjects can ask up to 20 draws per wheel. Each draw takes a little more than 3 seconds. When a subject decides to stop asking draws, (s)he has to estimate the probability that the next outcome of a wheel of fortune will be blue. In the flat fee treatment subjects earn the same amount irrespective of their estimates,

while subjects in the QSR treatment are paid according to the QSR. The same payment table was used as in experiment 1 (see appendix).

Results Experiment 2

Four subjects were removed from the analyses, because they apparently misunderstood the instructions. In the questionnaires these subjects revealed the false belief that every draw *within* a period was from *another* wheel of fortune or that the *same* wheel of fortune was used in *all* periods.[5] Of the remaining subjects, 23 participated in the QSR treatment and 33 in the flat fee treatment.

First, we look for extreme modes of behavior, like always reporting 50% or alternating between 0% and 100%. Only one subject in the flat fee treatment but none of the subjects in the QSR treatment always reported either 0% or 100%. In the QSR treatment, one subject reported 50% in 16 of the 20 periods and two subjects reported always 50% in the last 10 or 8 periods. One subject in the QSR treatment asked for two draws in the first period and for only one draw in all periods thereafter.[6]

The number of draws indicates the effort subjects are willing to invest in their task. One would expect that in the QSR treatment subjects ask for more draws. Figure 3 shows the number of draws for each period. First note that the number of draws asked by subjects in both treatments is remarkably high. In the first periods, subjects in the QSR treatment ask for more draws than subjects in the flat fee treatment, but thereafter differences rapidly disappear. The differences are not statistically significant (not overall, and not in the first periods only, Mann Whitney tests with individual averages as observations).

[5] The four subjects that are removed from the analyses are all from the QSR treatment. In the QSR-treatment the instructions were necessarily longer and more complicated than in the flat fee treatment.

[6] We have no indications that this subject misunderstood the instructions. He reported reasonable expectations: on average 36.25% if the draw was yellow and 60% if the draw was blue (a Bayesian would report 33% and 67% respectively).

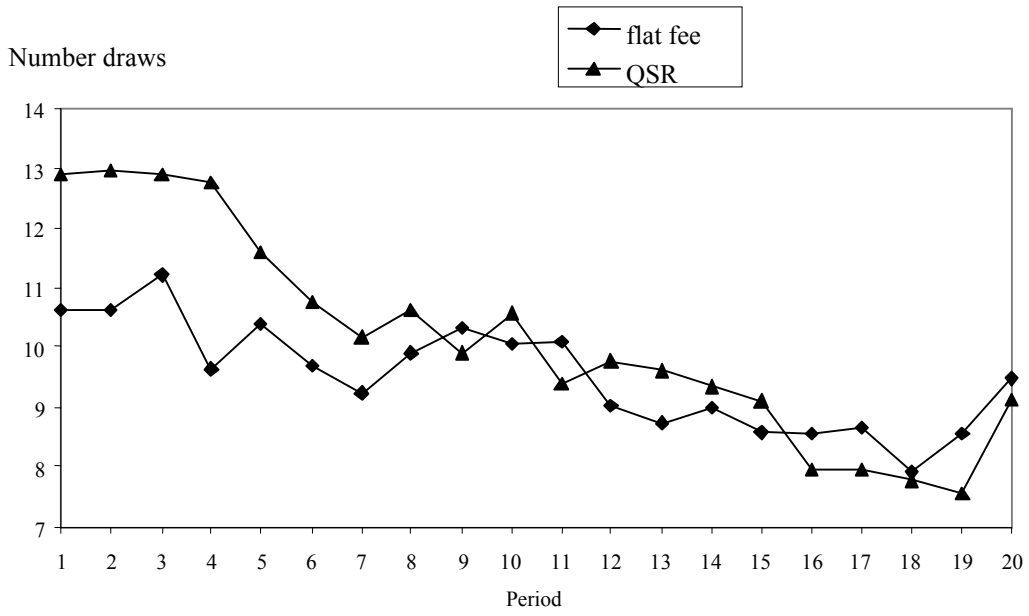

Figure 3: Number of draws per period for both treatments.

Even if treatments do not differ with respect to the effort subjects exert on searching for information, there could be a difference when they report their beliefs. In particular, one may conjecture that subjects in the flat fee treatment are more inclined to report only rough estimates. We have two standards to evaluate the reported expectations: Bayes rule and the (incorrect) proportional rule. First, we compare the beliefs of the subjects to the probabilities a Bayesian would have reported, based upon the same draws. If a subject observes n draws of which k are blue (denoted as n_k) the posterior distribution of the percentage blue of the wheel is:[7]

$$\Pr(B|n_k) = \frac{\Pr(B)\Pr(n_k|B)}{\Pr(n_k)} = \frac{\Pr(B)\Pr(n_k|B)}{\sum_{X=0}^{100}\Pr(X)\Pr(n_k|X)} = \frac{\frac{1}{101}\binom{n}{k}\left(\frac{X}{100}\right)^k\left(1-\frac{X}{100}\right)^{n-k}}{\sum_{X=0}^{100}\frac{1}{101}\binom{n}{k}\left(\frac{X}{100}\right)^k\left(1-\frac{X}{100}\right)^{n-k}} =$$

$$\frac{\left(\frac{X}{100}\right)^k\left(1-\frac{X}{100}\right)^{n-k}}{\sum_{X=0}^{100}\left(\frac{X}{100}\right)^k\left(1-\frac{X}{100}\right)^{n-k}} \qquad (2)$$

[7] The implicit assumption here is that the number of draws is independent of the outcomes of the draws.

The expectation of this distribution is:

$$\frac{\sum_{X=0}^{100} X\left(\frac{X}{100}\right)^{k}\left(1-\frac{X}{100}\right)^{n-k}}{\sum_{X=0}^{100}\left(\frac{X}{100}\right)^{k}\left(1-\frac{X}{100}\right)^{n-k}} \qquad (3)$$

For example, when 3 draws are observed, two yellow and one blue, the mean of the posterior distribution is 40%. However, subjects who neglect the base rate and use the proportion of blue draws as estimate, will report only 33%. In both treatments, the beliefs reported by most subjects are somewhat better organized by the (incorrect) proportional rule than to Bayes rule: 67% (61%) of the reported beliefs are closer to the proportional rule than to Bayes rule in the flat fee treatment (QSR treatment).[8]

The average (absolute) deviations of the reported expectations from the Bayesian estimates do not differ between the treatments, and neither do the average (absolute) deviations of the reported expectations from the proportional estimates.

Finally, the combined effect of the number of draws (more draws will give a better estimate) and the precision of the expectation formation process are compared between the two treatments. Figure 4 shows the real values and average expectations in all periods. Generally, subjects seem to do quite well. In a few periods in which the real probability is relatively extreme (periods 7, 16, 17 and 18), subjects seem to do less well in QSR than in the flat fee treatment, but the differences are not statistically significant.

In the experiment, the earnings in the QSR treatment were determined by one additional turn of each wheel (after the last period). To reduce noise we study the expected earnings: $Pr_{real}*(200*Pr_{exp}- Pr_{exp}^2)+(1- Pr_{real})*(10000- Pr_{exp}^2)$, with Pr_{real} the real probability of blue and Pr_{exp} the reported expectation of the subject. We also calculate the expected earnings for the subjects in the flat fee treatment (the expected payoff if they would have been in the QSR treatment). Figure 5 shows the results. In most periods we do not see any difference. The exceptions are again periods 7, 16, 17 and 18. No statistical significant difference is observed when all periods are pooled, but in periods 7, 16 and 18 expected earnings are less in QSR than in the flat fee treatment.

[8] In the questionnaire the subjects were presented with several sequences of outcomes and were asked their estimates. One of the sequences was Yellow-Yellow-Blue. 34 subjects (59%) filled in 30% or 33% and only 6 subjects (11%) filled in the bayesian 40%.

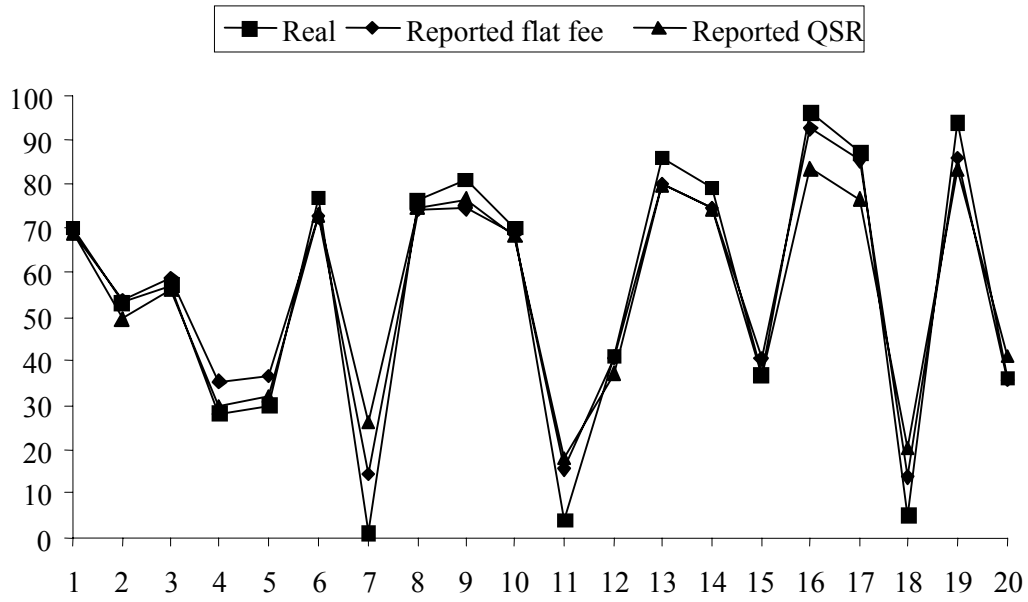

Figure 4: The real probability and the average of the reported probabilities for each period.

Figure 5: Expected earnings in the QSR treatment and the expected earnings in the flat fee treatment if subjects would have been paid according to the QSR.

4. Conclusions

The Quadratic Scoring Rule is theoretically a good instrument to elicit beliefs provided that respondents are risk neutral and do not distort probabilities. It has been criticized for these assumptions. In the first experiment, we presented a way to correct reported beliefs for the possible biasing effects of risk attitudes and probability weighting. We found that in practice these factors do not affect subjects' reported beliefs in an undesired way. This is reassuring news for previous studies that made use of QSR procedures without a correction device. However, in principle it is always possible to supplement the QSR procedure with the correction procedure described in the second part of experiment 1. In each setting there exists a possibility to map reported probabilities into true subjective probabilities. Although on the basis of the results of our experiment we would not expect much difference between reported and true subjective probabilities, it may be more prudent to supplement the QSR procedure with the correction procedure in future studies.

In the second study, the motivational aspects of the QSR treatment were studied. To our surprise, we found that subjects who were paid a flat fee exerted an equal amount of effort and were equally successful in estimating probabilities as subjects who were paid according to a QSR.

Sometimes experimental economists are very sceptical about the value of data that are generated without salient incentives. This scepticism is not supported by the results of our particular experiment. In fact, subjects do very well both when they receive salient incentives and when they do not. Given that subjects already performed so well without incentives, there was, of course, hardly any possibility for a positive effect of incentives.

We still believe that rewarding subjects for reporting their beliefs is the preferred procedure. For one thing, the task subjects face can make a difference. Maybe subjects enjoy the task in our experiment and therefore do not need any monetary incentive. Subjects have committed themselves to spend the time in the laboratory and try to do their best as long as their task is not too boring. Quite possibly the results will be different in situations in which subjects have to divide their attention between two tasks, e.g. a decision making task and a judgmental task. If the judgmental task is not rewarded, subjects may prefer to pay more attention to the salient decision making task. This possibility may be an interesting topic for future study.

References

Barron, Greg and Ido Erev (2000). On the Relationship between Decisions in One-Shot and Repeated Tasks: Experimental Results and the Possibility of General Models. Working paper, Technion, Haifa, Israel.

Beach, L.R. and L.D. Philips (1967). Subjective Probabilities Inferred from Estimates and Bets. *Journal of Experimental Psychology* 75, 354-359.

Camerer, Colin F. (1995). Individual Decision Making. In Kagel and Roth (eds.): *Handbook of Experimental Economics* 589-703.

Davis, Douglas D. and Charles A. Holt (1993) *Experimental Economics*. New Jersey: Princeton University Press.

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

Holt, Charles A. (1986). Scoring-Rule Procedures for Eliciting Subjective Probability and Utility Functions. In Goel, Prem K. and Arnold Zellner (eds.): *Bayesian Inference and Decision Techniques –Essays in Honor of Bruno de Finetti.* Amsterdam: Elsevier Science Publishers.

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

Jensen, F.A. and C.R. Peterson (1973). Psychological Effects of Proper Scoring Rules. *Organizational Behavior and Human Performances* 9, 307-317.

Kahneman, Daniel and Amos Tversky (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica* 47, 263-291.

Kraemer, Carlo, and Martin Weber (2001). To Buy or Not to Buy: Why Do People Buy Too Much Information? Working paper.

McDaniel, Tanga and Elisabet Rutström (2001). Decision Making Costs and Problem Solving Performance. *Experimental Economics* 4, 145-161.

McKelvey, R.D., and T. Page (1990) Public and Private Information: An Experimental Study of Information Pooling. *Econometrica* 58, 1321-1339.

Murphy, A.H. and R.L. Winkler (1970). Scoring Rules in Probability Assessment and Evaluation. *Acta Psychologica* 34, 273-286.

Nelson, R.G. and D.A. Bessler (1989). Subjective Probabilities and Proper Scoring Rules: Experimental Evidence. *American Journal of Agricultural Economics* 71, 363-369.

Nyarko, Yaw and Andrew Schotter (2000). An Experimental Study of Belief Learning with Elicited Beliefs. Working paper.

Offerman, Theo (1997). *Beliefs and Decision Rules in Public Good Games. Theory and Experiments*. Dordrecht Boston London: Kluwer Academic Publishers

Offerman, Theo and Joep Sonnemans (1998). Learning by Experience and Learning by Imitating Successful Others. *Journal of Economic Behavior and Organization* 34, 559-575

Offerman, Theo and Joep Sonnemans (2000). What is Causing Overreaction? An Experimental Investigation of Recency and the Hot Hand Effect. Working paper, downloadable at http://www.fee.uva.nl/creed/pdffiles/hothand.pdf

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

Offerman, Theo, Joep Sonnemans and Arthur Schram (2001). Belief Learning in Public Good Games. *Economic Inquiry* 39, 250-269.

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

Selten, Reinhard, Abdolkarim Sadrieh and Klaus Abbink (1999). Money Does Not Indue Risk Neutral Behavior, But Binary Lotteries Do Even Worse. *Theory and Decision* 46, 211-249.

Sonnemans, Joep, Arthur Schram and Theo Offerman (1998). Public good provision and public bad prevention. *Journal of Economic Behavior and Organization* 34, 143-161.

Sonnemans, Joep, Arthur Schram and Theo Offerman (1999). Strategic Behavior in Public Good Games: When Partners Drift Apart. *Economics Letters* 62, 35-41.

Appendix

Payoff table

Reported probability	Payoff if outcome is		Reported probability	Payoff if outcome is	
BLUE	BLUE	YELLOW	BLUE	BLUE	YELLOW
0%	0	10000	51%	7599	7399
1%	199	9999	52%	7696	7296
2%	396	9996	53%	7791	7191
3%	591	9991	54%	7884	7084
4%	784	9984	55%	7975	6975
5%	975	9975	56%	8064	6864
6%	1164	9964	57%	8151	6751
7%	1351	9951	58%	8236	6636
8%	1536	9936	59%	8319	6519
9%	1719	9919	60%	8400	6400
10%	1900	9900	61%	8479	6279
11%	2079	9879	62%	8556	6156
12%	2256	9856	63%	8631	6031
13%	2431	9831	64%	8704	5904
14%	2604	9804	65%	8775	5775
15%	2775	9775	66%	8844	5644
16%	2944	9744	67%	8911	5511
17%	3111	9711	68%	8976	5376
18%	3276	9676	69%	9039	5239
19%	3439	9639	70%	9100	5100
20%	3600	9600	71%	9159	4959
21%	3759	9559	72%	9216	4816
22%	3916	9516	73%	9271	4671
23%	4071	9471	74%	9324	4524
24%	4224	9424	75%	9375	4375
25%	4375	9375	76%	9424	4224
26%	4524	9324	77%	9471	4071
27%	4671	9271	78%	9516	3916
28%	4816	9216	79%	9559	3759
29%	4959	9159	80%	9600	3600
30%	5100	9100	81%	9639	3439
31%	5239	9039	82%	9676	3276
32%	5376	8976	83%	9711	3111
33%	5511	8911	84%	9744	2944
34%	5644	8844	85%	9775	2775
35%	5775	8775	86%	9804	2604
36%	5904	8704	87%	9831	2431
37%	6031	8631	88%	9856	2256
38%	6156	8556	89%	9879	2079
39%	6279	8479	90%	9900	1900
40%	6400	8400	91%	9919	1719
41%	6519	8319	92%	9936	1536
42%	6636	8236	93%	9951	1351
43%	6751	8151	94%	9964	1164
44%	6864	8064	95%	9975	975
45%	6975	7975	96%	9984	784
46%	7084	7884	97%	9991	591
47%	7191	7791	98%	9996	396
48%	7296	7696	99%	9999	199
49%	7399	7599	100%	10000	0
50%	7500	7500			

The numbers in the table are experimental francs. These francs will be exchanged after the experiment, 10000 franc equals 1.25 guilders.