



**HAL**  
open science

## Subjective beliefs formation and elicitation rules : experimental evidence

Guillaume Hollard, Sébastien Massoni, Jean-Christophe Vergnaud

► **To cite this version:**

Guillaume Hollard, Sébastien Massoni, Jean-Christophe Vergnaud. Subjective beliefs formation and elicitation rules : experimental evidence. 2010. halshs-00543828

**HAL Id: halshs-00543828**

**<https://shs.hal.science/halshs-00543828>**

Submitted on 6 Dec 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# Documents de Travail du Centre d'Économie de la Sorbonne

C  
E  
S  
W  
o  
r  
k  
i  
n  
g  
P  
a  
p  
e  
r  
s



## **Subjective beliefs formation and elicitation rules : experimental evidence**

Guillaume HOLLARD, Sébastien MASSONI, Jean-Christophe VERGNAUD

**2010.88**



# Subjective beliefs formation and elicitation rules: experimental evidence.\*

Guillaume Hollard<sup>1</sup>, Sébastien Massoni<sup>1†</sup>, and Jean-Christophe Vergnaud<sup>1</sup>

<sup>1</sup>CES, Université Paris 1

November 12, 2010

---

\*The authors are grateful to Karim N'Diaye, Thibault Gajdos and Peter Wakker; participants to ESE Conference in Rotterdam, EBIM Workshop in Paris, LabSi Workshop in Siena, FUR XIV in Newcastle, ESA 2010 in Copenhagen, SABE 2010 in San Diego for insightful comments. Financial supports from ANR grants is acknowledged (Riskemotion - ANR-08-RISKMAT-007-01; Feeling of control - BLAN07-2-192879). A previous version circulated with the title *Comparing three elicitation rules: the case of confidence in own performance*

†Corresponding author: [sebastien.massoni@gmail.com](mailto:sebastien.massoni@gmail.com)

## Abstract

Since they have been increasingly used in economics, elicitation rules for subjective beliefs are under scrutiny. In this paper, we propose an experimental design to compare the performance of such rules. Contrary to previous works in which elicited beliefs are compared to an objective benchmark, we consider a pure subjective belief framework (confidence in own performance in a cognitive task and a perceptual task). The performances of elicitation rules are assessed according to the accuracy of stated beliefs in predicting success. For the perceptual task we also compare stated beliefs to Signal Detection Theory predictions. We find consistent evidence in favor of the Lottery Rule which provides more accurate beliefs and is not sensitive to risk aversion. Furthermore the Free Rule, a simple rule with no incentives, elicits relevant beliefs and even outperforms the Quadratic Scoring Rule. Beside this comparison, we propose a belief formation model where we distinguish between two stages in the beliefs: beliefs for decision making and confidence beliefs. Our results give support to this model.

**Keywords:** Belief Elicitation, Confidence, Signal Detection Theory, Methodology, Incentives, Experimental Economics

**JEL Classification:** D81 D84 C60 C91

## Résumé

Depuis que leur utilisation s'est répandue en économie, les règles d'élicitation font l'objet d'une attention particulière. Dans ce papier, nous proposons une procédure expérimentale pour comparer les performances de telles règles. Contrairement aux travaux précédents dans lesquels les croyances élicitées sont comparées à des probabilités objectives, nous considérons ici un cadre de croyances purement subjectives (la confiance en sa propre performance dans une tâche cognitive et une tâche perceptive). Les performances des règles sont jugées en fonction de la capacité des croyances élicitées à prédire la réussite. Pour la tâche perceptive, nous comparons aussi les croyances élicitées aux prédictions issues de la détection de signal. Nos résultats sont en faveur de la *Lottery Rule* qui élicite des croyances plus justes et qui n'est pas dépendante de l'aversion au risque. De plus, la *Free Rule*, une simple élicitation sans incitation, élicite des croyances pertinentes et offre même de meilleurs résultats que la *Quadratic Scoring Rule*. Au-delà de cette comparaison, nous proposons un modèle de formation des croyances au sein duquel on distingue deux niveaux de croyances : celle utilisées pour la décision et celles liées à la confiance. Nos résultats supportent ce modèle.

**Mots clés :** Elicitation de croyances, confiance, Théorie de la détection de signal, Méthodologie, Incitations, Economie expérimentale

# 1 Introduction

Suppose that agents have somewhere in their mind subjective beliefs about some uncertain event. The history of beliefs' elicitation is rich as researchers have long been interested in this issue. In particular, following Cooke (1906) pathbreaking contribution, meteorologists have formally investigated this question for more than a century<sup>1</sup>. Quite surprisingly, it is only recently that economists have seriously investigated this issue. The current number of works focusing on empirical elicitation of beliefs has been rapidly increasing. For instance, Nyarko and Schotter (2002) study beliefs in experimental games, whereas Dominitz and Manski (1997) and Manski (2004) used surveys in order to elicit individual beliefs concerning significant personal events. There also exist a parallel literature in psychology. In many experiments subjects are asked to report their belief that they adequately performed a task. These particular type of beliefs are thus often referred to as confidence judgment or as metacognitive ability.

A common feature of these approaches is that they try to elicit "true" beliefs. This common goal however hides some differences across approaches. Meteorologists are mostly interested in *comparing* beliefs across individuals. They had thus been looking for elicitation rules that shape beliefs so as to express them on a common scale. Following the revealed preference approach, economists are interested in creating choice situations in which beliefs can be *inferred through costly actions*. They are thus very attached to elicitation rules that provide monetary incentives, so that subjects will lose money if they do not report their true belief. As the most popular elicitation rule in meteorology, namely the Quadratic Scoring Rule (QSR), does provide the kind of incentives economists are looking for, it has been imported in economics and has remained the most popular elicitation rule in experimental economics (see Gneiting and Raftery (2007) for a survey on proper scoring rules). Note that, so far neither economists nor meteorologists paid much attention to the type of reasoning that individuals may use to form their beliefs. In contrast, psychological experiments in which beliefs are elicited try to understand how the brain works when forming beliefs, or more precisely, when subjects are asked to provide judgment on their own performance. The main concern is to understand the *cognitive process associated with the formation of beliefs* (Dawes (1980), Baranski and Petrusic (1994)). Furthermore, the common practice is to use simple ordinal scales (e.g. Likert scales), without incentives. In what follows, we call this type of rule, "free" rules.

Thus, despite a common interest in beliefs elicitation, existing approaches still rest on

---

<sup>1</sup>See Murphy (1998) for an history of the early developments of beliefs elicitation

different elicitation rules. Does one of these approaches perform better than the others? To tackle this question, we propose to compare the relative performance of elicitation rules in an experimental setting in which subjects perform two tasks. The first one is a quiz task quite common in economics and psychology to assess subjects overconfidence (Lichtenstein and Fischhoff (1977), Lichtenstein, Fischhoff, and Phillips (1982), Wallsten and Budescu (1983) or Camerer and Lovo (1999), among others), the other task is a classical perception task often used to measure confidence. The most widely used elicitation rules are the Quadratic Scoring Rule -in economics and meteorology- and simple ordinal rules -in psychology-. However, there are growing methodological concerns about elicitation rules<sup>2</sup>. In particular, the QSR was found to have some important limitations. Thus we will consider a third elicitation rule into the analysis, called the Lottery Rule. We also propose a set of criteria allowing for elicitation rules to be compared. We find consistent evidence that the Lottery Rule outperforms both the QSR and the Free Rule whatever the nature of the task to be performed.

At a more fundamental level, this paper also offers insights on the nature of belief formation. Economists have often been concerned about the realism of their hypotheses and this also applies in the case of subjective probabilities. Axiomatic decision theory just argues that agents behave as if they maximize a subjective expected utility. Whether these subjective beliefs do really exist -in the sense that they can be directly measured using some physical device- remains an open question. However, Signal Detection Theory, a widely used approach in neurophysiology, offers a simple model of how the brain processes information. Applying this model to our data allows predicting confidence levels. We find that a particular elicitation rule, the Lottery Rule, elicits precisely the kind of signals that are assumed to be used in forming beliefs. So everything goes on as if the brain processes some signals that encode the beliefs and that a particular elicitation rule is able to elicit these signals. If such results are to be confirmed by further works, economists may well be able to open the black box of belief formation.

This article is organized as follows: in Section 2 we introduce our experimental devices with the psychometric and epistemic tasks and we set forth our beliefs formation models; in Section 3 we describe the main properties and the experimental design of the Quadratic Scoring Rule, the Lottery Rule and the Free Rule; in Section 4 we present the criteria to compare elicitation rules; in Section 5, we use the data to perform this comparison;

---

<sup>2</sup>See Andersen, Fountain, Harrison, and Rutstrom (2010), Offerman, Sonnemans, Van de Kuilen, and Wakker (2009), Palfrey and Wang (2009), Armentier and Treich (2010), Hao and Houser (2010), Hossain and Okui (2010) or Kothiyal, Spinu, and Wakker (2010)

in Section 6 we carry some further analysis and we suggest some interpretation of our results; Section 7 concludes.

## 2 Experimental devices for subjective beliefs

If agents have somewhere in their mind subjective beliefs that we try to elicit, then obviously, the closer the elicited beliefs are to the "true" one, the better is the rule. The main problem is that the "true" belief is in general not observable. Before presenting our approach, we first explain why we do not consider the case in which an *objective* probability is given. Until now most evaluation methods proposed have followed this latter methodology. For instance, one elicits a subject's beliefs concerning the possible outcome of a dice roll. The quality of the rule is then measured by the difference between the objective probability and the elicited beliefs. It is, for instance, the approach that Armentier and Treich (2010) have followed to assess the QSR using various level of stakes, either real or hypothetical. In a similar vein, Holt and Smith (2009) evaluate an explicit and incentive scoring rule (the so-called Lottery Rule) in an experiment on Bayesian updating. Note that these contributions do not compare different elicitation rules. So far as we know, the only contributions doing so are Hossain and Okui (2010), who observe that the rule they developed performs as well as the QSR, and Hao and Houser (2010) who compare two variants of the Lottery Rule. Both contributions take objective probabilities as a benchmark.

However, eliciting beliefs about objective probabilities might be misleading. First of all, even in a situation of objective uncertainty, the result will depend on how well subjects understand probability theory. If we assume that individuals deviate from the theory of probabilities, then their beliefs could differ from the objective probabilities. In such a case, a rule that allows for a perfect elicitation of individuals beliefs will be considered to perform poorly. Moreover, an objective probability setting may force subjects to think in terms of probabilities whereas they proceed differently in a subjective setting. In other words, the quality of an elicitation rule in a subjective setting cannot be deduced from its performance in an objective setting.

We thus face the following question: how the quality of elicited beliefs can be assessed without knowing the subjective beliefs we try to elicit? One contribution in this paper consists in answering this question. The key consists in submitting subjects to two successive exercises, i.e. take a decision and then assess their confidence in the decision made



through an elicitation rule. These sequences are repeated several times and through two experimental settings: a psychometric one and an epistemic one.

## 2.1 Psychometric model

Let us first consider the perceptual task, often used in psychophysics (Dawes (1980), Baranski and Petrusic (1994)). The aim of this task is to compare the number of dots contained in two circles (see Figure 2). The two circles are only displayed for a short fraction of time, about one second, so that it is not really possible to count the dots. Subjects have to tell which circle contains a higher number of dots and then, their confidence in the choice made is elicited through an elicitation rule.

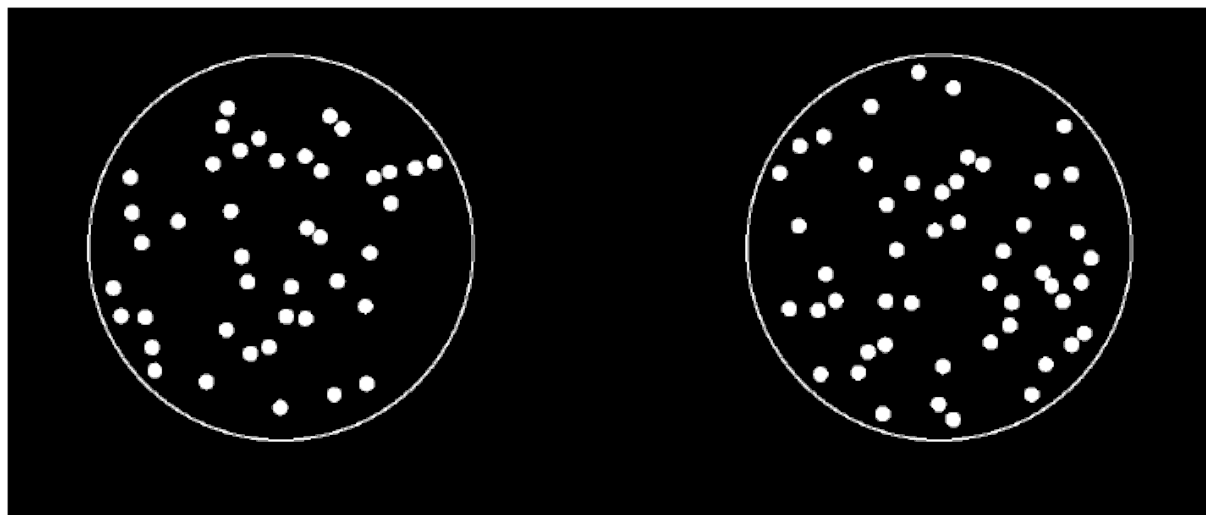


Figure 1: Perceptual task

We consider five levels of difficulty, i.e. bigger or smaller differences in the number of dots in each circle. The difficulty of the task depends on the subject's performance and is calibrated so that at each level the success rate is the same for each subject, using a psychophysics staircase (Levitt (1971)). Compared to the quiz questions, this setting offers to control for the difficulty of the task according to individual skills. Furthermore, as perceptual tasks are fast, it allows for a high number of trials.

A crucial feature of the perception setting is that it provides enough information to *predict* confidence level. To do so, we use Signal Detection Theory. The starting point of Signal Detection Theory is that most reasoning and decision making take place in the presence of some uncertainty (Green and Swets (1966)). According to Signal Detection Theory, subjects are assumed to receive a noisy signal provided by the sensory system

which treats the stimuli, e.g. in our task subjects only get some noisy information about the number of dots in each circle. In its most basic version, the model assumes that if  $x$  is the number of dots, the vision system sends a quantitative signal  $y$  which follows a normal law. Therefore, when observing two circles with respectively  $x_L$  and  $x_R$  dots (where  $L$  and  $R$  stand for left and right), the brain receives two signals  $y_L$  and  $y_R$ . Given the real difference  $\tilde{x} = x_L - x_R$ , the brain receives a  $\tilde{y} = y_L - y_R$  difference in signal which follows a normal law  $\mathcal{N}(x_L - x_R, \sigma_i^2)$  where  $\sigma_i$  reflects the sensibility quality of agent  $i$ 's vision system. Signal Detection Theory assumes that the brain system operates as if he was able to run Bayesian analysis of perceptive signals. Thus, to make his guess, the subject computes the posterior probabilities about the real difference  $\tilde{x}$  given the signal  $\tilde{y}$  received: he guesses left in case  $Pr(\tilde{x} \geq 0 | \tilde{y}) \geq .5$ .

We use this approach to predict the expected distribution of confidence<sup>3</sup>. The idea is the following (see the appendix for more detail): given the signal received  $\tilde{y}$ , the agent forms beliefs about the real difference in dots using Bayes law and choose accordingly. To apply Bayes law, we assume that the agent is aware of the probability distribution of dots used during the task and of his own sensibility quality  $\sigma_i$ <sup>4</sup>. We suppose then, that confidence in one's guess is closely related to the belief formed to make that decision. Thus, from the distribution of signal  $\tilde{y}$  given a certain  $\tilde{x}$  we can estimate a distribution of confidence  $p$  for each possible level of difficulty. Using the probability distribution of dots, expected distributions of confidence can be predicted as well as the values of the different criteria presented hereafter.

Figure 2 offers a representation of the way we assume the brain to process information. One area is responsible for receiving the signal that is transmitted to a distinct

---

<sup>3</sup>See Galvin, Podd, Drga, and Whitmore (2003) and Fleming and Dolan (2010) for a similar approach.

<sup>4</sup>In the task, there was always a circle which contains 50 dots and the second circle contains  $50 \pm \alpha_j$ . The choice of the 50 dots circle and of  $\alpha_j$  was randomized at each trial. In such a case, posteriors are such that a subject guesses left if  $\tilde{y}$  is positive. Five levels of difficulty were defined. Two were the same for all subjects: 1) level  $\alpha_0 = 0$ : the two circles contain 50 dots and success was randomly drawn, 2) level  $\alpha_4 = 25$ : the second circle contains 75 dots and the difference of dots becomes so important that this level leads to a sure choice. Three levels were intermediary and adapted to each subject. During the training part, the medium difficulty level  $\alpha_2$  was adjusted to a value in order to make the subject succeed in 70% of the case at this level. That means that  $\alpha_2$  is such that  $F(0 | \alpha_2, \sigma_i^2) = .3$  where  $F$  is the cumulative distribution for the normal law. The table of values indicates that  $\sigma_i = \frac{\alpha_2}{.52}$ . The two other levels were fixed respectively at  $\alpha_1 = \frac{\alpha_2}{2}$ . and  $\alpha_3 = 2\alpha_2$ . Then predicted success rate at level  $\alpha_1$  is  $1 - F(0 | \frac{\alpha_2}{2}, (\frac{\alpha_2}{.52})^2) \approx 0.60$  and  $1 - F(0 | 2\alpha_2, (\frac{\alpha_2}{.52})^2) \approx 0.85$  at level  $\alpha_3$ . In fact, the training part was not perfect and during the main task the mean success rate for all subjects was in reality at 67.7% at level  $\alpha_2$ . Then the model predicts that we should observe a 59% success rates at level  $\alpha_1$  and 82 % at level  $\alpha_3$ . Compared to the observed success rates which stand at 59% and 80% respectively, this model appears to be quite robust.

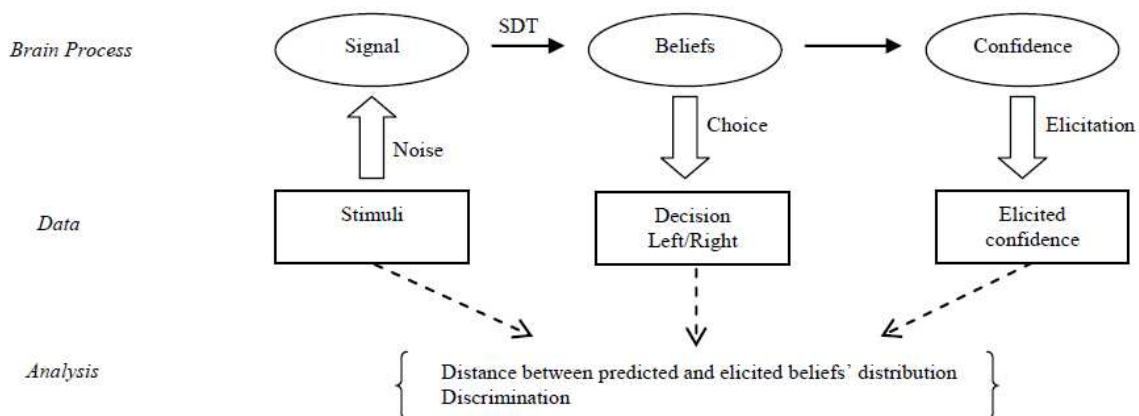


Figure 2: The psychometric model (perceptual task)

area in charge of forming beliefs. This signal is in turn used to assess confidence. Although very simple this representation of the brain is supported by recent evidence in neuroeconomics. For instance, Fleming, Weil, Nagy, Dolan, and Rees (2010) and Rounis, Maniscalco, Rothwell, Passingham, and Lau (2010) provide evidence that the brain areas responsible for performing the task and assessing confidence are distinct. We also know that signal transmission from one area to another could be problematic. Del Cul, Baillet, and Dehaene (2007) indeed show that subjects may use subliminal signal to perform adequately a task, i.e. they perform above chance, but are also unable to report accurate beliefs, i.e. their beliefs are not distinct from pure guessing. We interpret this finding as evidence that the brain adds some noise to the initial signals. Hence, our confidence distributions' predictions based on the SDT model are correct only if the noise is small. An extremely noisy case would be a person whose confidence is completely disconnected from beliefs.

Given this belief formation model, let us now explain why the choice of an elicitation rule may matter. It may be the case that elicitation rules do not elicit the same beliefs. They may elicit the beliefs formed at the decision stage or the confidence or something else... An other problem may be that rules are more or less efficient in revealing confidence signals. Indeed, it may depends on the physical type of these signals.

But then, how can we assess the relative performances of rules? First, even if the noise occurring in the brain process between beliefs and confidence makes our SDT predictions uncertain, if we find a good fit between elicited confidence and expected confidence this brings support to the related rule. Second, the SDT model shows that subjects are accu-

rate probability assessors of their success. Indeed, Bayes law implies that their beliefs are equal to the real probability of success. Thus, we should observe in the data that subjects are accurate probability assessors. Therefore, the subjects abilities to discriminate will be our main criteria to compare the rules.

## 2.2 Epistemic model

In our epistemic design, subjects have to answer a quiz of general knowledge and logic with "yes" or "no" answers. After each question, they have to give their level of confidence about the accuracy of their response. This level of confidence is elicited by three different eliciting rules presented in the next section. The main mechanism of the model could be summarized by the Figure 3.

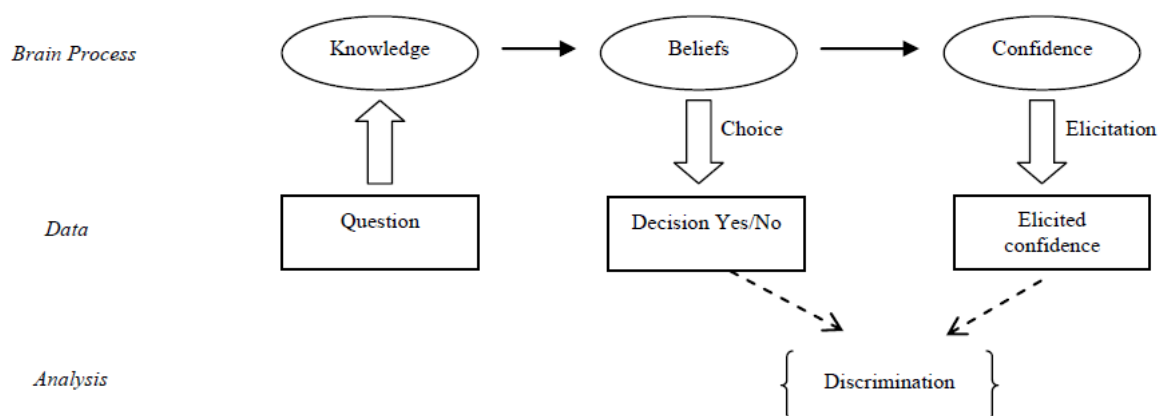


Figure 3: The epistemic model (quiz task)

We assume that, facing the quiz, the subject that takes his own level of knowledge as a basis, has a subjective probability about the answer to the question ("Yes" or "No"). This subjective belief determines his decision. Then we elicit the confidence the subject has in his answer. This two-steps protocol, often used in psychology, is representative of real situations. We often have to make a choice between two alternatives. Our subjective confidence in that choice determines the extent to which we commit to a path and we hedge our bets. Asking how much one believe the right answer is "Yes" allows for a direct elicitation of subjective beliefs but this way of proceeding is not very natural.

Even if the epistemic and the psychometric tasks imply different brain areas, we assume that the belief formation follows the same process. Thus, Figure 3 presents an adaptation of the previous model to the epistemic task. Contrary to the perceptive task where we

can rely on Signal Detection Theory, we do not know how beliefs correlate with success. Yet we suppose that subjects are still good probability assessors. Indeed, we expect that the more robust is one opinion, the more confident he is in his answer and the more likely he is to be right. Thus the quality of judgment of the subjects is still a relevant criterion for comparing rules.

Note that if the brain uses the same circuit to form beliefs in both perceptual and epistemic tasks, then we should observe some relation between tasks. Those who are good probability assessors in one task, should also be good in doing so in the other task. Such results would support this belief formation model.

### 3 Elicitation rules

In this section, we will describe the three types of rule used, discuss their main theoretical properties and present their experimental design.

#### 3.1 Quadratic Scoring Rule

In experimental economics, the most commonly used rule is the Quadratic Scoring Rule<sup>5</sup>. In its most simple version, when only two outcomes "success" or "failure" are considered, the Quadratic Scoring Rule rest on a score (or reward) of  $S_{success} = 1 - P_{failure}^2$  if "success" is the true state of nature and  $S_{failure} = 1 - P_{success}^2$  if "failure" is the true state of nature<sup>6</sup>. Remark that this standard theoretical presentation refers heavily to some subjective probabilities a subject has in mind. As noted in the introduction, this explicit reference is not necessary listed in the instructions provided to the subject.

Note also that this rule guarantees a sure payment when subjects report the same probability for each possible outcome. Thus risk averse subjects may prefer a sure payment

---

<sup>5</sup>Nyarko and Schotter (2002), Offerman, Sonnemans, Van de Kuilen, and Wakker (2009), or Palfrey and Wang (2009))

<sup>6</sup>Note that under this rule, a subject who reports a value of .5 for both probabilities will get a sure score of 0.5. Suppose now that he reports a probability of success of .7, he thus gets a gamble with .7 chance to get 0.91 and .3 chance to get 0.51. This results in an expected gain of 0.79. This extends to more general cases where there are  $n$  possible outcomes

$$S_i(p) = \alpha - \beta \sum_{k=1}^n (I_{i,k} - p_k)^2$$

were  $I_{i,k}$  takes value 1 if  $i = k$  and 0 elsewhere.

rather than a risky one. On a theoretical ground, it is well known that elicited beliefs through QSR for risk averse subjects are below their subjective beliefs<sup>7</sup>.

In our experiment we use the Quadratic Scoring Rule given in Table 1.

Choice										
Correct	10	9.98	9.90	9.78	9.60	9.38	9.10	8.78	8.40	7.98
Incorrect	0	0.98	1.90	2.78	3.60	4.38	5.10	5.78	6.40	6.98
7.5	6.98	6.40	5.78	5.10	4.38	3.60	2.78	1.90	0.98	0
7.5	7.98	8.40	8.78	9.10	9.38	9.60	9.78	9.90	9.98	10

Table 1: Quadratic Scoring Rule

Subjects can thus get a sure payment of 7.5€ or take greater risks, e.g. receive 10€ if their choice is correct, but 0€ if they fail. Note that the corresponding probability were not reported. Indeed, subjects were not told that if their confidence were at a certain probability level, then they should choose a particular column. We feel that this unusual presentation is more in line with a revealed preference approach and reduces confusion with the Free Rule.

### 3.2 Lottery Rule

In this experimental study, in order to elicit the level of confidence, we also use a procedure (henceforth, the *Lottery Rule*) whose principle is known for long (Arrow (1951), Raiffa (1968), Winkler (1972), LaValle (1978) among others...) but rarely put in practice (Grether (1992), Abdellaoui, Vossman, and Weber (2005), Holt (2006), Holt and Smith (2009) are some exceptions). Subjects are asked to report the beliefs about a given event, say their probability of success in a given task. Now consider a first lottery, called the task-lottery. According to the task lottery, the subject gets a positive reward,  $S$ , if he succeeds and a smaller reward  $F < S$ , if he fails. If his subjective probability of success is  $p$ , the subject should be willing to exchange his task lottery for any lottery that provides a reward of  $S$  with probability  $q > p$  (and reward  $F$  with probability  $1 - q$ ). Let us now consider the following mechanism: after the subject has reported a probability  $p$ , a random number  $q$  is drawn. If  $q$  is smaller than  $p$ , the subject is paid according to the

<sup>7</sup>Nevertheless, recent papers try to correct the QSR from risk attitudes (Offerman, Sonnemans, Van de Kuilen, and Wakker (2009), Andersen, Fountain, Harrison, and Rutstrom (2010), Kothiyal, Spinu, and Wakker (2010)).

task lottery. If  $q$  is greater than  $p$ , the subjects is paid according to a new lottery that provides the same reward with probability  $q$ , called the bonus lottery.

The Lottery Rule is very much in the line with the Becker-DeGroot-Marshak (1964) mechanism and does provide incentives to truthfully reveal  $p$ . To make this clear, suppose that the subject thinks the probability is  $p$  but reports a lower probability  $r$ . If the randomly chosen  $q$  is lower than  $r$ , the subject is paid according to the task lottery. If  $q > p$ , the subject benefits from the exchange of the lottery task to the bonus lottery. The interesting case arises when  $r < q < p$ . The subject is thus paid according to the bonus lottery, that has a *lower* probability of winning than the task lottery. So, the subject is worse off. Therefore underreporting  $p$  makes the subject worse off with a positive probability. Now consider a subject who overreports by stating a  $r$  above  $p$ . Again, the interesting case arises when  $r > q > p$ . In such a situation, the subject will not benefit from the bonus lottery and end up with the task lottery that has a lower probability of winning. So overreporting leads to an expected loss. Hence, the subject has then incentives to truthfully report his best estimates.

The true advantage of the Lottery Rule is that incentives are provided regardless of subjects' risk aversion<sup>8</sup>. Nevertheless, its main problem is that this rule is quite complicated and thus cognitively demanding. It is then of a particular interest to test whether the complexity of the Lottery Rule is indeed a problem.

In the experimental design, the Lottery Rule is implemented using a 0 to 100 scale, with steps of 5 (see Figure 4). The subjects received detailed explanations about the mechanism. The objective probability is determined using a uniform distribution between 40 and 100. Note that the nature of the distribution does not have any impact on the nature of provided incentives. Subjects receive 10€ for a correct answer if they are paid on the basis of the task lottery. If they benefited from the bonus lottery, a random draw determine whether they win. Remark that in such a case, their payment does not depend any longer on the quality of their answer. A favorable draw also leads to a payment of 10€.

### 3.3 Free Rule

The Free Rule just requires the subjects to report their beliefs, without relating any monetary consequences to stated probabilities. Nothing is done to provide incentives.

---

<sup>8</sup>More details and a formalization can be found in Karni (2009). Nevertheless Kadane and Winkler (1988) argue that this elicitation rule may not permit beliefs to be disentangled from utilities if the agents' wealth is correlated with the event. For the tasks we consider, this problem does not hold.

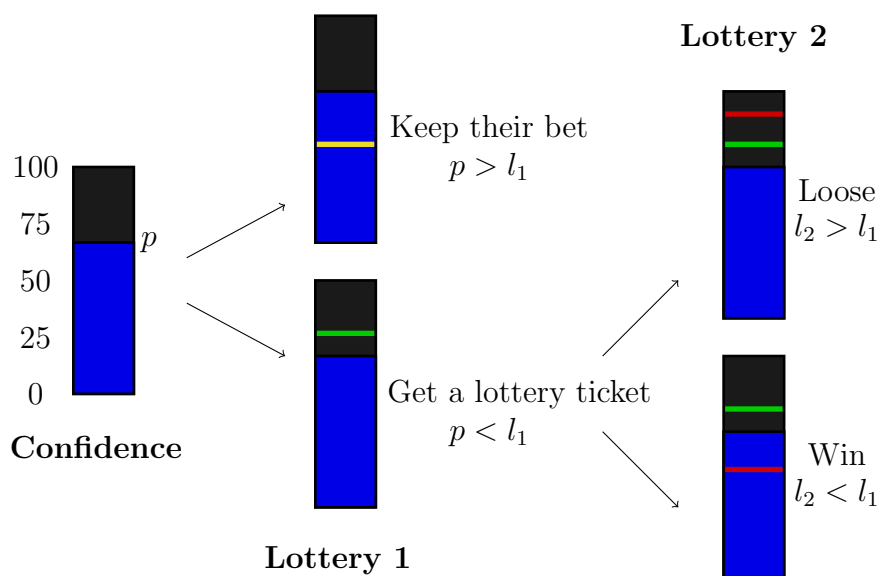


Figure 4: Lottery Rule

The strong advantage of such a rule is of course its simplicity. It is the less cognitively demanding one, especially comparing to the two previous ones.

The Free Rule is widely used in psychology and neurosciences. In particular, experiments that involve scanning the subjects are very sensitive to response times, as the duration of the experiment is limited and requires a high number of trials to obtain statistically significant results. Thus, the Free Rule is particularly attractive as beliefs are elicited in a very short period of time. It is also the case that psychologists are much less concerned by incentives than economists. So providing incentives for beliefs elicitation sometimes seems pointless, especially if incentives come at the price of a higher complexity.

Under the Free Rule (see Figure 5), the subject just has to choose a level of confidence between 0 and 100 (with steps of 5). Payments are only based on responses, whatever the accuracy of elicited beliefs. A correct answer to the selected quiz question provides a payment of 10€ (0 if incorrect).

## 4 Methods for comparing elicitation rules

In this section, we first present the three statistical tools we use to measure the quality of judgment and to compare elicitation rules and then we introduce the last criterion



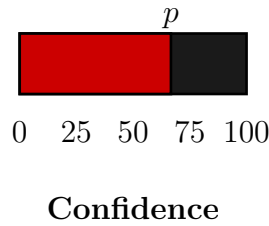


Figure 5: Free Rule

based on Signal Detection Theory. Even if one does not agree with the model proposed in section 2, comparing rules on the basis of the quality of probability assessment is still meaningful on a pure statistical basis. It is always better to use rules that allow for accurate forecasting.

#### 4.1 Calibration

To measure the accuracy of judgment, the most commonly used criterion is the distance between the mean predicted success rate and the actual one. This is the so called calibration criterion. Well calibrated stated beliefs are those which, on average, exhibit a small distance between predicted and actual success rate. The measure of calibration is relatively straightforward. Consider a subject who stated beliefs about  $n$  events,  $p_i$ , being his stated probability for event  $E_i$ ,  $x_i$  being the indicator variable that takes value 1 if he accurately predicts event  $E_i$ .

$$\text{calibration index} = \frac{1}{n} \sum_{i=1}^n (p_i - x_i)$$

A null value indicates that the subject is perfectly calibrated. A positive calibration indicates that the subject is overconfident, while a negative one denotes underconfidence. Note that by construction, the confidence predicted by the Signal Detection Model exhibits a perfect calibration  $\tilde{p}$ .

As we are interested in self-confidence related beliefs, we expect to find some overconfidence in our data. It is thus likely that individuals predict higher success rate than the one they really obtain. Indeed, it is common to consider that overconfidence and thus, miscalibration is a distinctive trait of many people (Camerer and Lovo (1999), Biais, Hilton, Mazurier, and Pouget (2005), Blavatsky (2009), Clark and Friesen (2009)). Asking for correct calibration is thus questionable.

## 4.2 Discrimination

Another important criterion to compare elicitation rules is discrimination. The ability to discriminate refers to the capacity of individuals to make a distinction between the probability of occurrence of two events. A subject that provides the same probability whatever the events will have a very low discrimination ability. It is important to note that such a subject might however be well calibrated if he reports for each trial a probability equals to his average success rate. Calibration and discrimination are thus two distinct notions, that both measure the accuracy of stated beliefs.

The corresponding statistical measure is given by the area under the ROC curve. Receiver Operating Characteristics (ROC) analysis is a graphical technique to visualize, organize and select classifiers according to their performance (Green and Swets (1966), Hanley and McNeil (1982))<sup>9</sup>. Here, a classifier is a dichotomous criteria based on a given level of confidence. Consider for example the classifier associated with the level of 0.7. This classifier will predict that each task that received a level of confidence higher than 0.7 will be classify as a success, while those with lower confidence will be classified as a failure. Such a classifier is not perfect. It sometimes predicts success when it should not, these are called the false positives. This allows the true positive rate to be computed (TPR), i.e. the fraction of predicted successes that are correctly predicted, and the false positive rate (FTP), i.e. the fraction of failures that are incorrectly predicted. Each classifier can then be represented on a two dimensional (TPR, FTP) space. Each level of confidence provides a point in this space. One can then fit a curve that relates these points, which is called the ROC curve. The area under the ROC curve (ROC Area) provides a measure of discrimination that is, the ability of the elicited confidence to correctly classify trials according to success or failure. To understand the meaning of the ROC area, we consider the situation in which trials are already correctly classified into two groups (success and failure) and we pick randomly a pair of trials, one from the success group and one from the failure group. The trial with the higher confidence should be the one from the success group. The area under the curve is the percentage of randomly drawn pairs for which this is true (that is, confidence correctly classifies trials in the random pair).

One advantage of the ROC analysis is that it uses confidence level only ordinally. For instance, if a subject is good at ranking his confidence but has some problem to give absolute values, his ROC Area can still be high. For our purposes, we prefer this latter criterion as it really catches the quality of the elicited confidence in terms of forecasting.

---

<sup>9</sup>See Kaivanto (2006) for an application in economics

### 4.3 Composite index

One important measure of overall performance in the accuracy of a judgment is the *Brier Score* (Brier (1950)). Following previous notation, the Brier Score is given by

$$BS = \frac{1}{n} \sum_{i=1}^n (p_i - x_i)^2.$$

The *Murphy Decomposition* (Murphy (1972), Yates (1982)) shows that the Brier Score aggregates a calibration and a discrimination index. Indeed, it can be expressed as:

$$BS = f(1 - f) - \frac{1}{n} \sum_{p \in P} N_p (f_p - f)^2 + \frac{1}{n} \sum_{p \in P} N_p (p - f_p)^2$$

where  $f = \sum_{i=1}^n (x_i)$  is the mean success rate,  $P$  is the set of possible probability judgments,  $N_p$  is the number of times that the confidence category  $p$  is used and  $f_p$  is the mean success rate in that class. The first term is the variance of the outcome variable which is independent of the judgment, the second term is a discrimination index measuring the ability of the hit rate around the overall base rate ( $f$ ) and the last term is a calibration index which measures the difference between the observed hit rate ( $f_p$ ) and the stated confidence. Note that contrary to the ROC analysis, confidence levels are used cardinally.

### 4.4 Distance between expected and elicited beliefs

The last criterion only holds in the psychometric setting. As we can predict using Signal Detection Theory the expected distribution of confidence for each subject, we can compare it to the elicited distribution of confidence. We compute a Chi-Square distance between the two distributions that allows for the comparison of two distributions. Formally the distance is computed as following:

$$\chi^2 = \sum_{p \in P} \frac{(N_p - E_p)^2}{E_p}$$

where  $E_p$  and  $N_p$  are respectively the expected and the observed number of the confidence level  $p$  occurrence. Then, for each rule we can compare these distances between observed and predicted distribution of confidence and identify which rule is the closest to the theoretical predictions.

## 5 Results

Before moving to comparative results, we will describe the main characteristics of the experiment and give some descriptive statistic stated confidences results.

### 5.1 Experimental design

The experiment took place in June and October 2009 at the Laboratory of Experimental Economics in Paris (LEEP). Subjects were recruited using LEEP's database. Most subjects were students from all fields. The experiments last for about 90 minutes. Subjects were paid 19 € on average. This computer-based experiment uses Matlab with the Psychophysics Toolbox version 3 (Brainard (1997)) and has been achieved on computers with 1024x768 screens.

Payments comprise three parts. The cognitive tasks are paid according to a standard procedure to avoid edging problems: one question is randomly selected at the end of the experiment and payments are computed according to the elicitation rule used. The case of the perceptual task is different, as each successful trial is rewarded according to the elicitation rule used (reward is 10cts for the Free and the Lottery Rules). Subjects also received a show-up fee of 5 €.

Our protocol enables us to investigate learning effects as we consider three sets of tasks. The same rule is used all along the session. The first block is composed of 36 quiz questions. Beliefs are elicited but no feedback is provided. Then subjects move on 100 trials of the perceptual task in which they get direct feedback (both on their accuracy and on their use of the rule). This should help subjects to improve their use of the elicitation rule. The third block is composed of 36 quiz questions, which are similar to the ones used in the first block, i.e. with no feedback. As we would like to compare the relative performance in the first and the third block, quiz questions were chosen so that they could be compared (similar subjects and similar success rates).

In each session, all beliefs were elicited using the same rule. We ran two session for each rule, that allowed to collect data for 35 to 38 subjects for each rule. As we mostly compare results between sessions, our design is a simple  $3 \times 1$  one.

### 5.2 Confidence data

As a preliminary step, we perform some descriptive analysis to draw a general picture of elicited beliefs. In Figure 6, we represent the cumulative probability distributions of

elicited confidence for both tasks and for each rule. The results for all subjects are pooled.

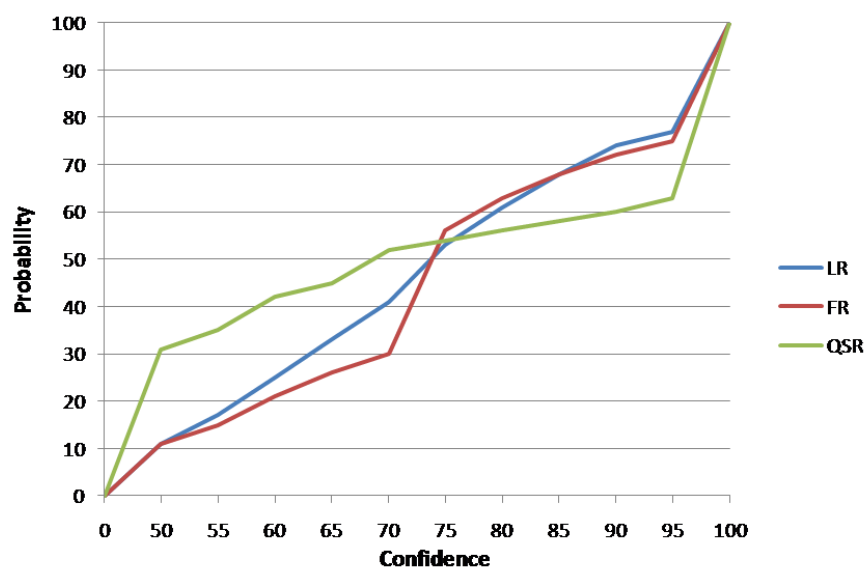


Figure 6: Cumulative probability distribution of confidence for each rule

It is clear that while the Lottery Rule and the Free Rule provide close cumulative distribution, the Quadratic Scoring Rule curve differs significantly from the two others. The difference is due to the fact that the QSR has a strong tendency to display stated probabilities concentrated on two values, 50% and 100%: almost two third of elicited probabilities range between these two values, this is twice as much as for the two other rules.

Let us now turn to a crude survey of how subjects, taken together, do judge themselves. Figure 7 compares the predicted success rate to the actual success rate, according to stated beliefs. A strong result, i.e. that applies for the three rules, is that subjects are globally overconfident. More precisely, the difference between expected and observed success rates increases for high level of stated confidence. Pooling all the tasks for which subjects stated a 100% probability of success leads to an actual success rate of about 78%. In contrast, low confidence (around 50%) leads to actual success rates that are roughly in line with expected ones. Even if the three rules are similar, some differences are worth being noted. None of the rules provides strictly increasing curves. It is not always the case that a 5% increase in stated probability leads to increase in the associated success rate. The most dramatic case is that of the QSR, for which there is not significant differences among stated probabilities in the range [65,95]. On average, any such probability leads to an

approximate rate of success of 67%.

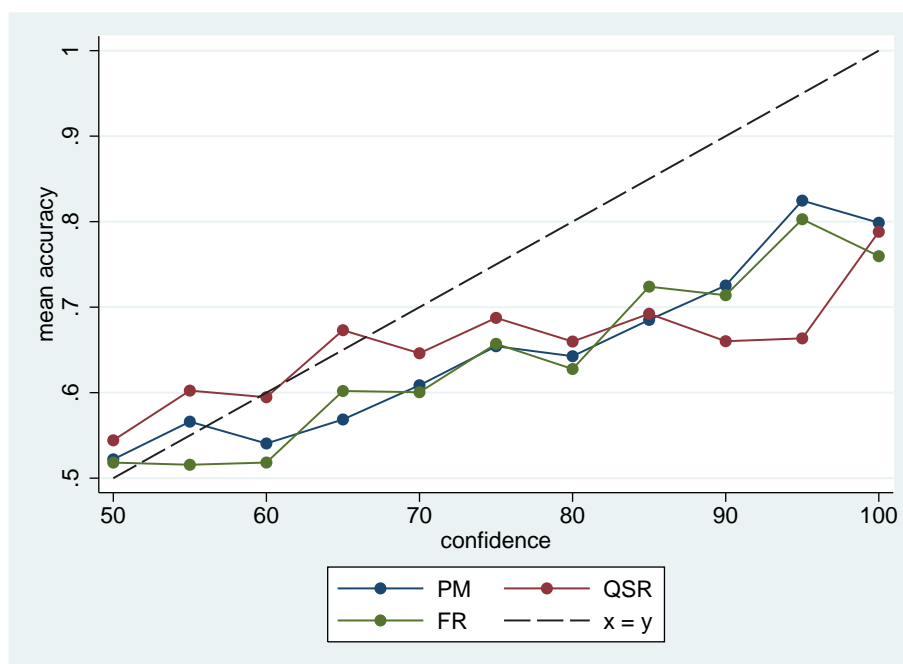


Figure 7: Matching between confidence and accuracy

This figure represents for the three rules the mean accuracy for each level of confidence between 50 and 100 with step of 5. We can see that the Lottery Rule and the Free Rule have a more regular and almost linear increasing function than the Quadratic Scoring Rule which takes on average a same level of accuracy for the intermediate level of confidence.

Conversely, we also observe the classical "hard-easy" effect (Lichtenstein and Fischhoff (1977)), that is, overconfidence for hard task and underconfidence in easy task. In Figure 8, we draw the mean success rate and confidence values for each question of the quiz. We clearly see on the left-hand side of the figure that high overconfidence is associated with the most difficult quiz questions and on the right, underconfidence with the less difficult questions. Note that some questions are misleading e.g. 80% of the subjects are pretty sure that they got the correct answer and thus stated high confidence, while in fact they were wrong. The Figure 8 provides some indication about the frequency of such questions. Removing these misleading questions will thus diminish overconfidence by almost half.

The "hard-easy" effect is also observed for the perceptual task. Note that the Signal Detection Model predicts such an effect. Indeed, since subjects form belief on the basis of a Bayesian analysis of noisy signals, they are overconfident when the difficulty is high ( $\alpha_0$  or  $\alpha_1$ ) as erroneous signals emerge and are underconfident when the difficulty is low ( $\alpha_3$  or  $\alpha_4$ ). In Table 2, we report the observed and predicted confidence as well as the observed success rate.

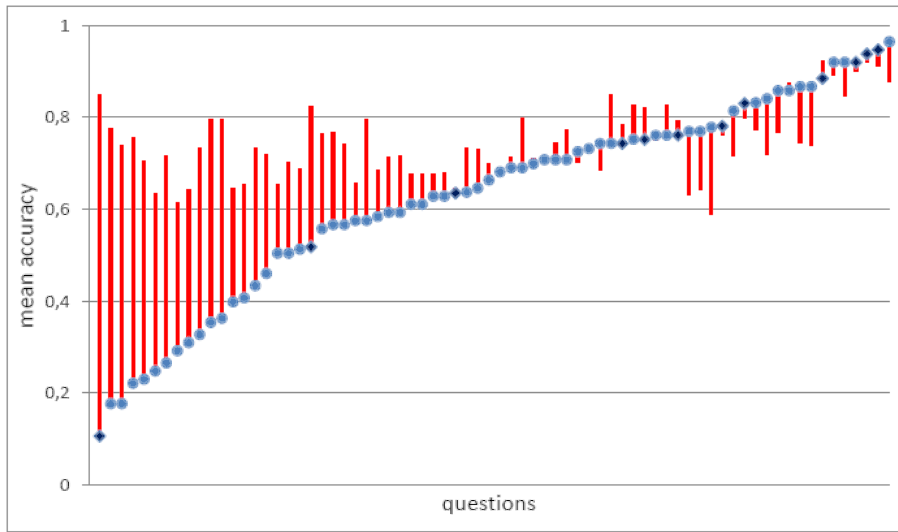


Figure 8: Ranked success rate and overconfidence to quiz questions

This figure shows the level of mean accuracy for each question of the quiz. The circle corresponds to general knowledge questions and the diamond to logical questions. The red bar is the level of over/under confidence depending on the direction (up/down) of this bar. As we can expect, we find an "hard-easy effect", i.e. a greater overconfidence for more difficult sets of questions and some underconfidence for easy questions.

	Observed mean confidence			Predicted mean confidence			Success rate		
	LR	QSR	FR	LR	QSR	FR	LR	QSR	FR
$\alpha_0$	71%	69%	72%	64%	63%	64%	45%	52%	50%
$\alpha_1$	72%	71%	74%	64%	64%	64%	60%	60%	58%
$\alpha_2$	72%	73%	74%	66%	65%	66%	69%	67%	67%
$\alpha_3$	77%	76%	78%	71%	69%	70%	82%	79%	80%
$\alpha_4$	94%	96%	94%	95%	95%	91%	100%	100%	100%

Table 2: Observed and predicted mean confidence. (Std. in brackets)

Notice that the Signal Detection Model predicts very small differences in mean confidence for the three more difficult levels while success rates range from around 50% ( $\alpha_0$ ) to 70% ( $\alpha_2$ ). Empirically, these predictions are confirmed for the three rules.

### 5.3 Quality of judgment

We turn now to the statistical index of quality judgment. Table 3 provides some measures using the statistical index described above. The QSR performs better in terms of calibration, as it displays a lower degree of overconfidence than the two other rules. This result is a weak support for QSR since it is plagued by risk aversion and since overconfidence

is well established. The QSR is likely to generate an underestimated overconfidence. As expected from Figure 7, the Lottery Rule provides a better discrimination than the QSR with an area under the ROC curve of 0.6401. Using the composite index, Lottery Rule clearly outperforms the two other rules with a Brier Score of 0.2245. The more striking results is that the Free Rule appears slightly better than the QSR.

Rule	Overconfidence	ROC Area	Brier Score
LR	0.0822 (.0057)	0.6401 (.0070)	0.2245
QSR	0.0668 (.0061)	0.6300 (.0073)	0.2262
FR	0.1065 (.0060)	0.6305 (.0074)	0.2259
(LR - QSR)	+0.0153 (0.0329)	+0.0101 (0.3186)	-0.0017 (0.0026)
(LR - FR)	-0.0243 (0.0017)	+0.0096 (0.3458)	-0.0014 (0.0314)
(QSR - FR)	-0.0396 (0.0000)	-0.0005 (0.9608)	+0.0003 (0.3830)

Table 3: Comparison of rules

This table (as the following) summarizes the values and the tests of differences of the three criteria used to evaluate the accuracy of confidence for the three rules. For each rule, we have the level of overconfidence with the standard deviation, the area under the ROC curve (with s.d.) and the value of the Brier Score. Then, we compare the rules by pairs and we find the level of difference for each criteria. We perform a test of difference: for the overconfidence and the Brier Score we test the significance of the inequality between the rules (t-test with the p-value in parenthesis); and for the ROC Area we test the significance of the equality between the rules (Chi-square test with p-value in parenthesis).

One may wonder whether this ranking is robust to learning effects. The QSR and the Lottery Rules are cognitively demanding and we expect their performance to increase with practice. Our experiments is designed so as to offer subjects the opportunity to learn using feedback. Remember that we use three blocks of questions. The second one relates to the perceptual task with feedback. The idea was to use this task as a training phase for the elicitation rule. Therefore, we can compare beliefs' accuracy in the first and third block (where the tasks to be performed are quiz questions of similar difficulty). The Table 4 provides details about the learning effect under the three different rules. It compares the relative performance in the two sets of quiz questions. The overall effect is limited and its direction is unclear. If learning occurs, we should observe more accurate results, i.e. better calibration or better discrimination. Our results does not support this view. The most significant effect, if any, is found for the Lottery Rule. The Brier Score increases from 0.2464 to 0.2548, mainly because calibration is not as good as in the first part. This does not completely rule out the possibility that subjects indeed learn as another effect, e.g. subjects get tired, might work in the opposite direction. But even if this is the case, we can conclude that none of the rules display a clear advantage in terms of learning.



Tasks	Brier Score	Overconfidence	ROC Area
LR_Q1	0.2464	0.1117 (.0131)	0.6158 (.0154)
LR_Q2	0.2548	0.1326 (.0130)	0.6318 (.0151)
LR_(Q2 - Q1)	+0.0084 (0.0000)	+0.0209 (0.1282)	+0.0160 (0.4594)
QSR_Q1	0.2607	0.0786 (.0142)	0.5770 (.0161)
QSR_Q2	0.2589	0.0934 (.0141)	0.5949 (.0157)
QSR_(Q2 - Q1)	-0.0018 (0.1842)	+0.0148 (0.2295)	+0.0179 (0.4270)
FR_Q1	0.2586	0.1510 (.0137)	0.6143 (.0158)
FR_Q2	0.2590	0.1448 (.0138)	0.5885 (.0161)
FR_(Q2 - Q1)	+0.0004 (0.4177)	-0.0062 (0.3741)	-0.0258 (0.2533)

Table 4: Learning: cognitive tasks

## 5.4 Perceptual data

The superiority of the Lottery Rule and the Free Rule over QSR is confirmed when we compare stated confidence to the Signal Detection Model predictions. Indeed, let us consider for the perceptual task the cumulative probability distributions for each rule and for the Signal Detection Model predictions (all subjects pooled) drawn in Figure 9. Contrary to the QSR, the shapes of the LR and the FR curves follow the Signal Detection Model.

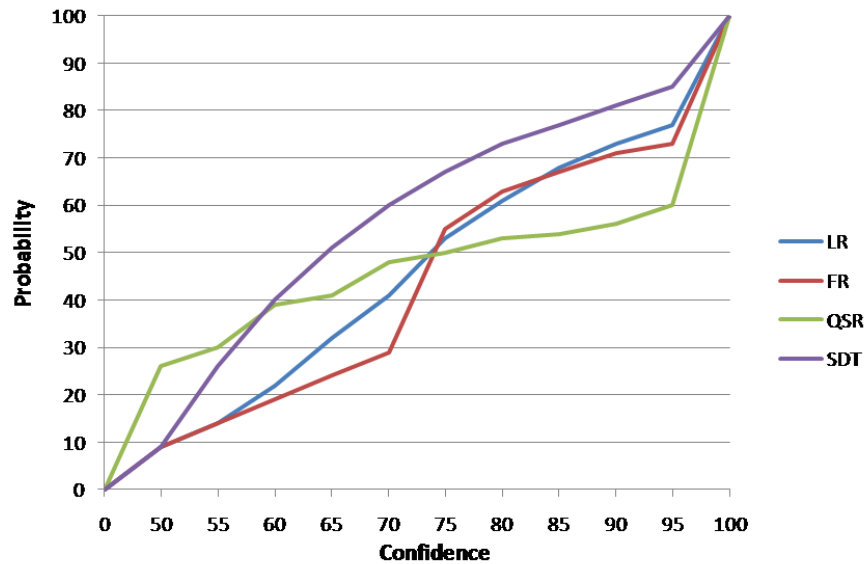


Figure 9: Cumulative probability distribution of confidence for each rule and SDT

The difference between the QSR curve and the three others is also due to the concentration of probabilities around 50% and 100% with again almost two third of elicited probabilities that take these values. Note that SDT predicts that only 24% of stated confidence should take these two values (see Table 12 in appendix)<sup>10</sup>. This visual feeling is confirmed when we compute the Chi-Square distance between the observed and predicted confidence distribution for each group of subjects. To get a fixed idea about the distance values, we also provide the distance between the predicted confidence and the uniform distribution on [50%; 100%] as well as the distance between the predicted confidence and a Dirac measure that puts a probability of 1 on 100%. Results are given in Table 5.

Chi-square distance btw. pred. confid. and...	LR (n=38)	QSR (n=35)	FR (n=35)
...observed confidence distribution	0.20	1.17	0.80
...a uniform distribution	0.29	0.31	0.28
...a Dirac measure $\delta_{100\%}$	5.29	5.97	6.08

Table 5: Chi-square distance between confidence distributions

The Lottery Rule clearly outperforms the QSR in its ability to fit with predicted confidence<sup>11</sup>. Further analysis confirms this result.

For the perceptual task, subjects go through a training phase during which an automatic adjustment of difficulty was done so as to make the subjects succeed at 70% for the medium difficulty level. This adjustment was not perfect and during the main task subjects' success rate differs with a overall mean success of 71.2% (67.7% for the medium level), a standard deviation of mean success equal to 6% and a mean success ranging from 56% to 86%. Accordingly, the Signal Detection Model predicts that mean confidence should be lower for those whose the task was more difficult given their abilities (see appendix for detail). Therefore, we should find a correlation between elicited and predicted mean confidence. In the Table 6, we give the observed and predicted mean confidence values as well as the correlation between observed and predicted mean confidence.

We observe a significant correlation between the observed and predicted mean confidence (at a level of 10%) only for the Lottery Rule while there is no correlation for the QSR.

---

<sup>10</sup>The percentage of elicited confidence concentrated on 50% and 100% is 27% for LR, 35% for FR and 65% for QSR

<sup>11</sup>Despite its weaknesses, the QSR performs decently in terms of discrimination because it permits to correctly classify easy trials (difficulty levels  $\alpha_3$  and  $\alpha_4$ ) and difficult trials (difficulty levels  $\alpha_0$  and  $\alpha_1$ ). We guess that QSR discrimination performance would be lower for a hard task without low levels of difficulty.

	Nb	Observed mean confidence	Predicted mean confidence	Correlation
LR	38	77% (9%)	72% (5%)	0.30*
QSR	35	77% (10%)	71% (4%)	0.02
FR	35	79% (9%)	71% (7%)	0.12

Table 6: Observed and predicted mean confidence. (Std. in brackets) (\* means significance at 10%)

The Signal Detection Model also predicts that those for which the task was more difficult given their ability should have a lower discrimination index, either measured by the ROC area or by the Brier score. Therefore, we expect to find correlations between mean success rates and discrimination indexes. Furthermore, we should also observe correlations between observed and predicted discrimination indexes. Clearly, we see in Table 7 that the levels of correlation are poorer for the QSR than for the two other rules. For instance, the predicted correlation between the mean success rate and ROC area for the LR group is 0.8439 and 0.4855 for the predicted and observed beliefs while these correlations are respectively 0.9082 and 0.0074 for the QSR group. It is even worse for correlation between predicted and observed ROC area as there is not any correlation at all for the QSR group. Among the two other rules, the Lottery Rule rule performs slightly better than the Free Rule.

Correlation		Nb	Mean succ./ROC	Mean succ./Brier	ROC obs/pre	Brier Obs/pre
LR	Pred.	38	0.8439***	-0.8728***	0.4951***	0.7825***
	Obs.		0.4855***	-0.8633***		
QSR	Pred.	35	0.9082***	-0.8700***	-0.0143	0.4272**
	Obs.		0.0074	-0.4253**		
Free	Pred.	35	0.9455***	-0.9336***	0.4617***	0.7680***
	Obs.		0.3498**	-0.7761***		

Table 7: Correlation between mean success rate, ROC Area and Brier score (observed and predicted) for the task perception (\*\*\* means a level of significance at 1%. \*\* means a level of significance at 5%)

## 6 Further results and interpretation

Existing work in psychology and neuroscience puts forward some regularities about confidence. If SDT happens to be a good model of the way subjects processes information,

there are a few additional predictions that can be tested. For instance, Fleming, Weil, Nagy, Dolan, and Rees (2010) and Rounis, Maniscalco, Rothwell, Passingham, and Lau (2010) provide evidence revealing that the brain areas responsible for performing the task and assessing confidence are distinct. Therefore the simple description of the brain process depicted on Figure 2 and 3 accounts for empirical evidence. Everything goes as if subjects receive a noisy signal that is first used to perform the task, then transmitted -up to some noise- to another area responsible for assessing confidence. If we follow the two models proposed in Figures 2 and 3, discrimination ability reflects how well the brain areas responsible for performing the task are connected to the brain areas responsible for assessing confidence.

For some subjects we do indeed observe that the mismatch occurs: subject unable to discriminate have their confidences disconnected from subjective beliefs. This raises the question of whether subjects who are good at discriminating on one task are also good when proceeding another task. There are some experimental evidence showing that people are overconfident over domains even if their levels of overconfidence may vary with the domain, and that more overconfident people in one domain tend also to be more overconfident in other domain (West and Stanovich (1997)). For discrimination, evidence are scarce but according to Bornstein and Zickafoose (1999), it seems also to be the case that discrimination abilities export across domains. In Table 8, we report our findings on the correlation between tasks for calibration and discrimination.

Corr. btw. Quiz and Perception	All (104)	LR (38)	QSR (34)	FR (32)
Calibration	0.57***	0.47***	0.69***	0.49***
ROC area	0.27***	0.35**	0.15	0.33*

Table 8: Correlation between tasks

For all rules, we find some high correlations for calibration. For discrimination, we find significant correlation only for the Lottery Rule and for the Free Rule. Note that SDT predicts that subjects cannot have a high ROC area if their task was hard in the perceptual task. Thus, to measure the subject intrinsic discrimination ability, we must take into account the predicted ROC area which stands for benchmark value. So we created a ROC performance attainment measure defined as follows:

$$ROC_{pa} = \frac{ROCArea(observed) - 0.5}{ROCArea(predicted) - 0.5}$$

We expect this variable to correlate with the ROC area observed for the quiz task. In Table 9 we report these correlations.

	All (104)	LR (38)	QSR (34)	FR (32)
Corr. btw. quiz ROC area and ROC <sub>pa</sub>	0.10	0.31*	0.00	0.07

Table 9: Correlation between ROC in Quiz and ROC performance in Perception.

Significant correlation remains only for the Lottery Rule. Thus discrimination ability seems more domain specific than calibration ability. Note also, that the choice of elicitation rule strongly matters. Indeed, the Lottery Rule results confirms our conjecture on intrinsic discrimination ability while the QSR results invalidates it.

Calibration and discrimination are two statistically independent aspect of judgment ability. We may wonder whether subjects who are more able on one dimension, do so in the other. Table 10 indicates correlations between calibration and discrimination for subjects whose overconfidence is below 30%<sup>12</sup>.

Corr. btw. Calibration and ROC area	All (104)	LR (38)	QSR (34)	FR (32)
All task	-0.14	-0.06	-0.35**	-0.01
Quizz task	-0.20**	-0.17	-0.39**	-0.18
Perception task	-0.11	-0.07	-0.14	-0.12

Table 10: Calibration versus discrimination

We find significant correlation only for the Quiz task and this arises from the result obtained using the QSR. Because of the poor overall performance of the QSR, we may conclude that calibration and discrimination abilities are independently distributed in the population.

To sum up, we find that the LR matches the prediction of SDT and also satisfies some additional properties. We interpret this as supporting the view that LR indeed elicits signals that are correlated with the ones uses by the brain to assess confidence. On the whole, our results supports the model set forth in Section 2. An intriguing question is then to figure out why the LR fits so well, while another rule like QSR does not. A possible interpretation is that for the LR -as well as the Free Rule- subjects are directly asked to report the beliefs they consciously have in mind. In contrast, the QSR requires

<sup>12</sup>We exclude outliers because they correspond to subjects who always choose extreme confidence and therefore were highly miscalibrated and discriminated poorly

the subjects to take the monetary rewards into account. They need to evaluate for each level of confidence how high they should bet. At the very least, subjects have to perform some additional computing to achieve a good decision. This extra computation is likely to introduce some additional noise into the signal we are trying to elicit. But we also learn from the post-wagering literature, that rules in which subjects are asked to set their wages themselves -like the QSR- are dependent on some economic variables, typically risk aversion. Reviewing current evidence, Fleming and Dolan (2010) conclude that "the complex interaction between objective stimulus visibility, wager size and the subsequent willingness to gamble casts doubt on the assertion (Persaud, McLeod, and Cowey (2007)) that post-decision wagering is a direct index of subjective awareness, despite its intuitive nature."

## 7 Conclusion

Both our experimental settings provide consistent evidence: the choice of a particular elicitation rule does matter. The Lottery Rule performs particularly well according to the discrimination index. The Lottery Rule also matches remarkably well the predictions made using Signal Detection Theory. Furthermore, the Lottery Rule has the theoretical advantage to be not sensitive to risk aversion. All in all, we find little support for the use of the QSR in economics. The Free Rule performed well. Although it elicits a bit less accurate beliefs, it is the simplest rule that can be implemented. Thus elicitation rules which ask subjects to report their feelings in terms of a visual metric outperform elicitation rules which are based on a revealed preference approach through the choice of stakes. The fact that incentives are not so important supports the view that there exist pure subjective beliefs that are disconnected from utility values.

Another interesting result is the high heterogeneity we find in individual's discrimination abilities. In our experiment discrimination ability -as measured by the ROC area- ranges from 0.51 to 0.79. In other words some subjects are almost unable to discriminate at all, while others perform remarkably well. However, subjects with low discrimination abilities perform as well as high discrimination abilities' subjects, i.e. their success rates are not statistically different. This important variability across individuals is confirmed by recent neuroeconomic findings (Fleming, Weil, Nagy, Dolan, and Rees (2010), Rounis, Maniscalco, Rothwell, Passingham, and Lau (2010)). They find that discrimination ability is linked to how well the brain areas responsible for performing the task are connected

with the brain areas responsible for assessing confidence. The ability to discriminate is also independent from their calibration ability. Formally, calibration and discrimination are statistically independent: one can be very well calibrated without discriminating well and vice versa. In economics, most research focus on overconfidence (Camerer and Lovo (1999), Biais, Hilton, Mazurier, and Pouget (2005), Blavatskyy (2009), Clark and Friesen (2009)). So far discrimination has not received much attention while it is certainly as important as calibration to explain economic behavior.

## 8 Appendix

### 8.1 Signal detection model

We detail here how we apply the Signal Detection Theory to induce subjective beliefs.

#### 8.1.1 The basis for the prediction of confidence distribution

Let us consider for instance a subject who receives a signal  $\tilde{y} = y_L - y_R > 0$ . The signal  $\tilde{y}$  follows a Normal Law with a mean equals to the real difference in dots  $\tilde{x} = x_L - x_R$  and a variance  $\sigma_i^2$  where  $\sigma_i$  reflects the sensibility quality of the subject. We assume that the subjects' brain is aware of the quality of his vision system and of the distribution of dots used during the task. We then apply a Bayesian analysis. First, notice that since the priors are symmetric between left and right, if the subject receives a positive signal he will believe with a probability above .5 that the real signal is positive, i.e  $P(\tilde{x} \geq 0 | \tilde{y} \geq 0) \geq .5$  and thus he guesses left. In the following, we suppose that the subject confidence in guessing right is equal to his belief.

Thus, given a value  $\tilde{y}$ , the subject confidence in winning is equal to

$$P(\tilde{y}) = \text{Proba}(\tilde{x} = x_L - x_R > 0 | \tilde{y}) + .5\text{Proba}(\tilde{x} = x_L - x_R = 0 | \tilde{y}).$$

The second term catches the probability of evenness between  $x_L$  and  $x_R$  so that the subject wins with a .5 probability. By Bayes law,

$$P(\tilde{y}) = \frac{\text{Proba}(\tilde{y} | \tilde{x} > 0) \cdot \text{Proba}(\tilde{x} > 0) + .5\text{Proba}(\tilde{y} | \tilde{x} = 0) \cdot \text{Proba}(\tilde{x} = 0)}{\text{Proba}(\tilde{y})}$$

Under the assumption that the brain is aware of the distribution of dots used during

the task, then:

$$\text{Proba}(\tilde{x} = 0) = \text{Proba}(\tilde{x} = \pm\alpha_i) = \text{Proba}(\tilde{x} = \pm 25) = .2$$

and thus

$$P(\tilde{y}) = \frac{\left( \sum_{j=0,\dots,4} f(\tilde{y}|\tilde{x} = \alpha_j) \right)}{\left( \sum_{j=0,\dots,4} f(\tilde{y}|\tilde{x} = \alpha_j) \right) + \left( \sum_{j=0,\dots,4} f(\tilde{y}|\tilde{x} = -\alpha_j) \right)}$$

with  $f$  the density function of the normal law.

Similar computation for a negative signal  $-\tilde{y}$  shows that  $P(-\tilde{y}) = P(\tilde{y})$ . We note that confidence  $P(\tilde{y})$  is strictly increasing in  $|\tilde{y}|$ . Then, the probability to observe a confidence level  $\tilde{p}$  is the probability that the brain receives a signal  $\tilde{y}$  such that  $P(|\tilde{y}|) = \tilde{p}$ . Given  $\alpha_j$ , the density function for the confidence  $\tilde{p}$  is equal to

$$g(\tilde{p}) = .5 \left( \begin{array}{l} f(\tilde{y}|\tilde{x} = \alpha_j) + f(-\tilde{y}|\tilde{x} = \alpha_j) + \\ f(\tilde{y}|\tilde{x} = -\alpha_j) + f(-\tilde{y}|\tilde{x} = -\alpha_j) \end{array} \right) \text{ for } \tilde{y} \geq 0 \text{ such that } P(|\tilde{y}|) = \tilde{p}$$

In the experiment, confidence was elicited with a path of 5%. We proceed similarly for the prediction of confidence. Hence, we suppose that an elicited confidence of 50 % corresponds to an underlying confidence between 50% and 52.5%, of 55 % corresponds to an underlying confidence between 52.5% and 57.5% and so on .....Therefore, given  $\alpha_j$ , the probability of observing  $p \in \{.55; .60; \dots\}$  is

$$Q(p|\alpha_j) = \int_{\tilde{p}=p-.025}^{\tilde{p}=p+.025} g(\tilde{p})d\tilde{p}.$$

and overall, the predicted distribution of confidence is given by

$$Q(p) = .2 \sum_{j=0,\dots,4} Q(p|\alpha_j).$$

By construction, the predicted confidence reflects perfect calibration, that is, the mean success rate is equal to  $\tilde{p}$  when confidence is  $\tilde{p}$ :

$$\text{Proba}(\text{Correct Guess}|\tilde{p}) = \text{Proba}(\tilde{x}.\tilde{y} > 0|\tilde{p}) + .5\text{Proba}(\tilde{x} = 0|\tilde{p}) = \tilde{p} \text{ for } \tilde{y} \text{ such that } P(|\tilde{y}|) = \tilde{p}.$$



For our estimates, we make the approximation that when pooling confidence, we also have perfect calibration, i.e:

$$\text{Proba}(\text{Correct Guess}|p) = p \text{ for } p \in \{.50; .55; .60; \dots; 1\}$$

### 8.1.2 Implementation details

The model is applied at an individual level. The first step is to estimate for each individual his ability  $\sigma_i$ . His ability is revealed through his success rate at levels  $\alpha_{j=1,\dots,4}$ . At level  $\alpha_j$ , we observe  $n_{i,j}$  trials and  $r_{i,j}$  successes. We compute  $\sigma_i$  such that

$$\sum_{j=1,\dots,4} n_{i,j} \cdot F(0|\alpha_j, \sigma_i^2) = \sum_{j=1,\dots,4} r_{i,j}$$

The following table gives the descriptive statistics for the  $\sigma_i$ .

	Nb	Mean	Std. Dev.	Min;Max
$\sigma_i$	108	7.95	5.30	2.8 ;50.5

Table 11: Ability of subjects

From  $\sigma_i$ , we can compute for each  $i$  and level  $\alpha_j$  the confidence distribution on  $p \in \{.55; .60; \dots\}$ :

$$Q_i(p|\alpha_j) = \int_{\tilde{p}=p-.025}^{\tilde{p}=p+.025} g(\tilde{p}) d\tilde{p}.$$

The overall confidence distribution is then computed using the observed levels' frequencies:

$$Q_i(p) = \sum_{j=0,\dots,4} \frac{n_{i,j}}{100} \cdot Q_i(p|\alpha_j).$$

Some descriptive statistics for the confidence distribution are given in the Figure 10.

Given these confidence distributions, we can calculate the predicted judgment quality index. Calibration is not really an issue since predicted mean confidence, predicted mean success and observed mean success should be very close by construction. The only divergence comes from the fact that empirical success rate at level  $\alpha_0$  may not be exactly 50% because of a low number of trials and of the approximation done in the estimation of confidence. We can check in the Table 12 that it is indeed the case.

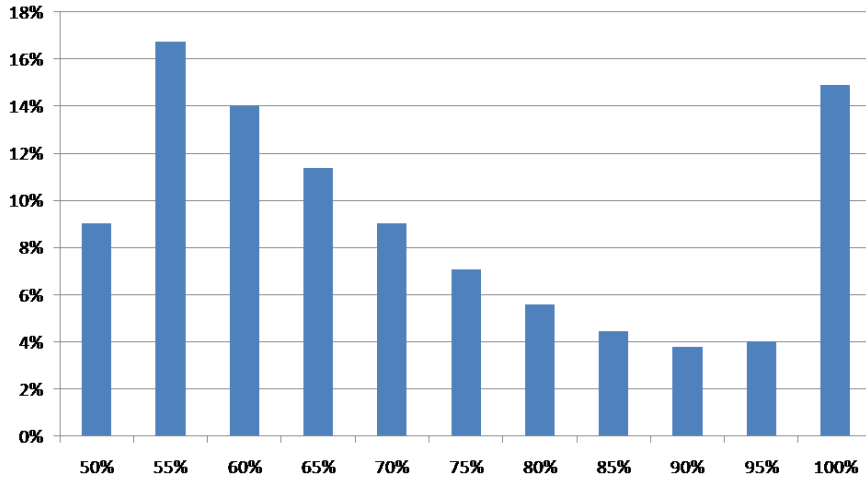


Figure 10: SDT prediction of confidence distribution

Difference btw...	Nb	Mean	Std. Dev.	Min;Max
predicted and observed mean success	108	.00	.02	-.04;.07
predicted mean success and predicted mean confidence	108	-.00	.01	-.02;.02

Table 12: Consistency of prediction

For discrimination, to calculate predicted area under the ROC curve we first estimate predicted True Positive Rate and False Positive Rate at each cutpoint. For instance, if we consider fix confidence level  $p$  as a cutpoint, TPR is given by:

$$\text{Proba}(\text{Confidence} > p | \text{Success}) = \frac{\text{Predicted mean success} - \sum_{p' \leq p} p' \cdot Q_i(p')}{\text{Predicted mean success}}$$

and FPR is defined by:

$$\text{Proba}(\text{Confidence} > p | \text{Failure}) = \frac{\sum_{p' > p} (1 - p') \cdot Q_i(p')}{\text{Predicted mean failure}}$$

Given estimation of TPR-FPR at each cutpoint  $p \in \{.50; .55; .60; \dots; 1\}$ , a predicted ROC Area can be computed for each subject.

Finally, the computation of predicted Brier Score by subject is done using the following formula:

$$\sum_{p \in \{.50; .55; .60; \dots; 1\}} Q_i(p) \cdot [p(p-1)^2 + (1-p) \cdot p^2]$$

The Table 13 gives a summary of the values obtained

	Nb	Mean	Std. Dev.	Min;Max
ROC Area	108	.72	.05	.55;.83
Brier score	108	.18	.03	.12;.25

Table 13: Predicted ROC area and Brier score

## References

- ABDELLAOUI, M., F. VOSSMANN, AND M. WEBER (2005): “Choice-Based Elicitation and Decomposition of Decision Weights for Gains and Losses under Uncertainty,” *Management Science*, 51(9), 1384–1399.
- ANDERSEN, S., J. FOUNTAIN, G. HARRISON, AND E. RUTSTROM (2010): “Estimating Subjective Probabilities,” *CEAR Working Paper*.
- ARMENTIER, O., AND N. TREICH (2010): “Eliciting Beliefs: Proper Scoring Rules, Incentives, Stakes and Hedging,” *Working Paper*.
- ARROW, K. J. (1951): “Alternative Approaches to the Theory of Choice in Risk-Taking Situations,” *Econometrica*, 19, 404–437.
- BARANSKI, J., AND W. PETRUSIC (1994): “The Calibration and resolution of confidence in perceptual judgments,” *Perception and Psychophysics*, 55(4), 412–428.
- BIAIS, B., D. HILTON, K. MAZURIER, AND S. POUGET (2005): “Judgmental Overconfidence, Self Monitoring, and Trading Performance in an Experimental Financial Market,” *The Review of Economic Studies*, 72(2), 287–312.
- BLAVATSKYY, P. (2009): “Betting on own knowledge: Experimental test of overconfidence,” *Journal of Risk and Uncertainty*, 38(1), 39–49.
- BORNSTEIN, B. H., AND D. J. ZICKAFOOSE (1999): “‘I Know I Know It, I Know I Saw It’: The Stability of the Confidence-Accuracy Relationship Across Domains,” *Journal of Experimental Psychology: Applied*, 5(1), 7688.
- BRAINARD, D. (1997): “The Psychophysics Toolbox,” *Spatial Vision*, 10, 433–436.

- BRIER, G. W. (1950): “Verification of Forecasts Expressed in Terms of Probability,” *Monthly Weather Review*, 78(1), 1–3.
- CAMERER, C., AND D. LOVALLO (1999): “Overconfidence and Excess Entry: An Experimental Approach,” *The American Economic Review*, 89(1), 306–318.
- CLARK, J., AND L. FRIESEN (2009): “Overconfidence in forecasts of Own Performance: An Experimental Study,” *The Economic Journal*, 119(534), 229–251.
- COOKE, E. (1906): “Forecasts and verifications in Western Australia,” *Monthly Weather Review*, 34(1), 23–24.
- DAWES, R. (1980): “Confidence in intellectual judgments versus confidence in perceptual judgments,” in *Similarity and choice: Papers in honor of Clyde Coombs*, ed. by E. Lantermann, and H. Feger, pp. 327–345. Han Huber.
- DEL CUL, A., S. BAILLET, AND S. DEHAENE (2007): “Brain Dynamics Underlying the Nonlinear Threshold for Access to Consciousness,” *PLoS Biology*, 5(10), e260.
- DOMINITZ, J., AND C. F. MANSKI (1997): “Using Expectations Data to Study Subjective Income Expectations,” *Journal of the American Statistical Association*, 92, 855–867.
- FLEMING, S. M., AND R. J. DOLAN (2010): “Effects of Loss Aversion on Post-decision Wagering: Implications for Measures of Awareness,” *Consciousness and Cognition*, 19(1), 352363.
- FLEMING, S. M., R. S. WEIL, Z. NAGY, R. J. DOLAN, AND G. REES (2010): “Relating Introspective Accuracy to Individual Differences in Brain Structure,” *Science*, 329, 1541–1543.
- GALVIN, S. J., J. V. PODD, V. DRGA, AND J. WHITMORE (2003): “Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions,” *Psychonomic Bulletin and Review*, 10, 843876.
- GNEITING, T., AND A. E. RAFTERY (2007): “Stricly Proper Scoring Rules, Prediction, and Estimation,” *Journal of the American Statistical Association*, 102(477), 359–378.
- GREEN, D. M., AND J. A. SWETS (1966): *Signal detection theory and psychophysics*. John Wiley and Sons.

- GREETHER, D. (1992): “Testing Bayes rule and the representativeness heuristic: Some experimental evidence,” *Journal of Economic Behavior and Organization*, 17, 31–57.
- HANLEY, J. A., AND B. J. MCNEIL (1982): “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” *Radiology*, 143, 29–36.
- HAO, L., AND D. HOUSER (2010): “Getting It Right the First Time: Belief Elicitation with Novice Participants,” *Working Paper*.
- HOLT, C. (2006): *Markets, Games, and Strategic Behavior: Recipes for Interactive Learning*. Addison-Wesley.
- HOLT, C., AND M. SMITH (2009): “An Update on Bayesian Updating,” *Journal of Economic Behavior and Organization*, 69(2), 125–134.
- HOSSAIN, T., AND R. OKUI (2010): “The Binarized Scoring Rule of Belief Elicitation,” *Working Paper*.
- KADANE, J. B., AND R. L. WINKLER (1988): “Separating Probability Elicitation From Utilities,” *Journal of the American Statistical Association*, 83(402), 357–363.
- KAIVANTO, K. (2006): “Informational rent, publicly known firm type, and ‘closeness’ in relationship finance,” *Economics Letters*, 91(3), 430–435.
- KARNI, E. (2009): “A Mechanism for eliciting Probabilities,” *Econometrica*, 77(2), 603–606.
- KOTHIYAL, A., V. SPINU, AND P. WAKKER (2010): “Comonotonic Proper Scoring Rules to Measure Ambiguity and Subjective Beliefs,” *Working Paper*.
- LAVALLE, I. H. (1978): *Fundamentals of Decision Analysis*. Holt, Rinehart and Winston, New York.
- LEVITT, H. (1971): “Transformed up-down methods in psychoacoustics,” *Journal of the Acoustical Society of America*, 49, 467–477.
- LICHTENSTEIN, S., AND B. FISCHHOFF (1977): “Do those who know more also know more about how much they know? The calibration of probability judgments,” *Organizational Behavior and Human Performance*, 20(7), 159–183.

- LICHTENSTEIN, S., B. FISCHHOFF, AND L. PHILLIPS (1982): “Calibration of probabilities: the state of the art to 1980,” in *Judgment under uncertainty: Heuristic and biases*, ed. by D. Kahneman, P. Slovic, and A. Tversky, pp. 306–334. Cambridge University Press.
- MANSKI, C. F. (2004): “Measuring Expectations,” *Econometrica*, 72(5), 1329–1376.
- MURPHY, A. H. (1972): “Scalar and Vector Partitions of the Probability Score. Part I: Two-State Situation,” *Journal of Applied Meteorology*, 11, 273–282.
- (1998): “The early history of probability forecasts: Some extensions and clarifications,” *Weather and Forecasting*, 13, 5–15.
- NYARKO, Y., AND A. SCHOTTER (2002): “An Experimental Study of Belief Learning Using Elicited Beliefs,” *Econometrica*, 70(3), 971–1005.
- OFFERMAN, T., J. SONNEMANS, G. VAN DE KUILEN, AND P. WAKKER (2009): “A Truth-Serum for Non-Bayesian: Correcting Proper Scoring Rules for Risk Attitudes,” *Review of Economic Studies*, 76(4), 1461–1489.
- PALFREY, T., AND S. WANG (2009): “On Eliciting Beliefs in Strategic Games,” *Journal of Economic Behavior and Organization*, 71(2), 98–109.
- PERSAUD, N., P. MCLEOD, AND A. COWEY (2007): “Post-decision wagering objectively measures awareness,” *Nature Neuroscience*, 10, 257–261.
- RAIFFA, H. (1968): *Decision Analysis*. Addison-Wesley, London.
- ROUNIS, E., B. MANISCALCO, J. C. ROTHWELL, R. E. PASSINGHAM, AND H. LAU (2010): “Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness,” *Cognitive Neuroscience*, 1(3), 165–175.
- WALLSTEN, T. S., AND D. V. BUDESCU (1983): “Encoding Subjective Probabilities: A Psychological and Psychometric Review,” *Management Science*, 29(2), 151–173.
- WEST, R. F., AND K. E. STANOVICH (1997): “The Domain Specificity and Generality of Overconfidence: Individual Differences in Performance Estimation Bias,” *Psychonomic Bulletin and Review*, 4(3), 387–392.
- WINKLER, R. L. (1972): *An Introduction to Bayesian Inference and Decision Theory*. Holt, Rinehart and Winston, New York.

YATES, J. F. (1982): "External Correspondence: Decompositions of the Mean Probability Score," *Organizational Behavior and Human Performance*, 30(1), 132–156.