

Comparing three elicitation rules: the case of confidence in own performance

Guillaume Hollard¹, Sébastien Massoni^{1*}, and Jean-Christophe Vergnaud¹

¹CES, Université Paris 1

June, 2010

PRELIMINARY DRAFT: Please do not cite or quote without permission.

Abstract

The aim of this paper is to compare three elicitation rules, the quadratic scoring rule, the matching probability rule and the free rule. To do so, we compare the behavior of three groups of subjects who are asked to perform the same tasks. The tasks to be performed are cognitive and perceptual tasks. We only vary the rule used to elicit their beliefs.

Our main finding is that the elicitation rule used does matter. According to various criteria, the probability matching rule provides more accurate beliefs. As this rule also has the advantage of not being sensitive to risk aversion, it has a clear advantage over the quadratic scoring rule. The free rule does a decent job at eliciting beliefs and even outperforms the quadratic scoring rule according to various criteria.

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

1 Introduction

Beliefs are key elements of decision theory. Eliciting beliefs in experiments is thus a major concern. Elicitation methods for encoding subjective beliefs are numerous and diverse. They notably differ on two aspects. First, there is the question of monetary incentives to report beliefs truthfully. Among the rules which are based on monetary incentives, proper scoring rules constitute a major group. Economists are attached to the revealed preference approach and therefore are much in favor of monetary incentives. The common practice in experimental economics is thus to use scoring rules which link payment of subjects to elicited beliefs. Among proper scoring rules, the Quadratic Scoring Rule (QSR) is the most popular one in economics. In psychology and neuroscience, this issue is less preeminent and beliefs data gathered without

*Corresponding author: sebastien.massoni@gmail.com

incitation are considered as valuable. A second aspect on which elicitation methods differ is how much their instructions refer explicitly to subjective degrees of belief. In this respect, one extreme method is to ask subjects to express their beliefs directly in terms of percentage of chance. On the other extreme, in the wagering method (Persaud and al (2007)) beliefs are inferred from the subjects' betting behavior. Usually, instructions for proper scoring rules refer to subjective probabilities but it could be possible to phrase the task of choosing the stakes without making any reference to subjective probabilities.

Do these discrepancies in elicitation methods matter? Are there rules performing better than others? It is not straightforward to know how to answer to these questions and a clarification of what it means to elicit subjective beliefs is first needed. It depends on whether we hold the view that agents have a set of internal coherent probabilities clearly stated in their mind or not. If we consider that for any event E , an agent holds a clear subjective probability $\pi(E)$, then a good elicitation rule is one which makes the agent report an elicited belief $p(E)$ close to $\pi(E)$. It is well known that elicitation rules based on monetary incentives and thus where the subjects have to choose the stakes, are subject to distortion (see Kadane and Winkler (1988), Offerman and alii (2009)). In particular, under the quadratic scoring rule a risk neutral subject should report his "true" belief. In other words, reporting anything else than the "true" belief leads the subjects to an expected loss. This incentive property comes at a price however. Risk averse subjects (and risk seeking ones) are expected to misreport their beliefs. The quadratic scoring rule is thus expected to distort elicited beliefs, but evidence on the importance of this effect are scarce in the literature so far. If one doubts about the "internal view" that an agent holds clear subjective beliefs, then the analysis becomes more complex. This is for instance the position of Wallsten and Budescu (1983):

"Rather, an individual's opinion is more-or-less vaguely formulated, and upon being asked to evaluate the probability of an outcome, a person will search his or her memory for relevant knowledge, combine it with the information at hand, and (presumably) provide the best judgment, possible. That judgment will depend on what is retrieved from memory, what aspects of the current information are utilized, and possibly on the sequential order in which this is all integrated into a unified opinion" (p153).

Furthermore, we may well envisage that the elicitation rule interfere in the process of making a judgment. At this stage, note that the revealed preference approach is of no aid. Indeed, it just says that, beliefs as well as utilities are not directly observed, but rather are inferred through their manifestations in behavior. Therefore, if one considers that subjective beliefs do not exist per se, then he should not expect that elicited beliefs through a proper scoring rule or through a simple verbal report should coincide. Then what is a good elicited $p(E)$ if we do not believe in the existence of $\pi(E)$? The revealed preference approach is not totally empirically vacuous: it presumes that elicited beliefs through different means should be the same. Consider for instance a quiz task. When asked, if a subject guesses that E is true, he revealed that $p(E) > p(\neg E)$. When we elicitate his confidence, we suppose that this confidence is equal to $p(E)$. Even if we doubt that there exist a precise subjective probability $\pi(E)$ driving both the guessing task and the confidence elicitation, it is realistic to think that the subject computes "evidence signals" $v(E)$ and $v(\neg E)$ and guesses on this basis. The hypothesis that confidence in E increases in $v(E)$ and that the higher is $v(E)$, the higher is the probability for the subject of guessing right seems natural. On the basis of this hypothesis, we should expect a correlation between confidence and success. It is already known that this correlation fails in some setting, for instance the case of "subliminal" stimuli where subjects perform above a chance level but report

zero confidence in their guessing (see Del Cul and al. (2007)). Indeed, “subliminal” stimuli receive a perceptual treatment and even permit to increase performance but these processes remain inaccessible to consciousness. While rather extreme, the “subliminal” stimuli example indicates that belief signals in the brain may be diverse and elicitation rules may be connected to more or less noisy belief signals. Thus, the idea is to judge the quality of an elicitation rule by the quality of the elicited beliefs. One facet of judgment quality is calibration, that is, the match between subjective probabilities with the real actual success rate. Discrimination is an other facet. 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 has very low discrimination ability will provide the same probability whatever the events. It is important to note that such a subject might however be well calibrated if he reports for each trial a probability equal to his average success rate. Discrimination seems to be the primary criteria to compare elicitation rules since discrimination measures the correlation between confidence and success. Calibration is not so important since miscalibration or overconfidence is considered to be a widespread bias and since some elicitation rules induce distortion. For instance, subject may appear to be well calibrated with the Quadratic Scoring Rule because their risk aversion compensates their overconfidence.

The aim of this paper is to compare three representative elicitation rules. One is the Quadratic Scoring Rule. We call the Free rule the simple rule where subjects are simply asked to report their beliefs on a given scale and no monetary consequences derive from reported beliefs. The main advantage of the Free rule is its simplicity. It is easy to handle and to explain. Furthermore, it saves some time as it takes subjects only a few second to report their beliefs. The temporal dimension is important in experiment which require a high number of trial like typical scan experiments. At a time when behavioral economics and neuroeconomics intend to strengthen the links between economics, psychology and neurosciences, it could be of interest to consider the Free rule and to compare its property to the rules used in experimental economics. The last rule is the Probability Matching rule, also called lotteries by Kadane and Winkler (1988), for which some economists has recently regained interest in (Holt (2006) Holt and Smith (2009), Hao and Houser (2010)). The basic principle is to offer the subject the opportunity to exchange a lottery based on his subjective probability against a lottery with identical payoff but known probability. Subjects are thus expected to accept such exchange as long as the objective probability is larger than their subjective one. The trick is to use a mechanism that provides incentives to reveal the truth by using a Becker-DeGroot-Marshak type of mechanism (see below for a detailed description). The Probability Matching rule have the desirable property that incentives are provided regardless of subjects’ risk aversion¹. But, again, there is a price to pay as the Probability Matching rule is quite complicated and thus cognitively demanding. It is then of particular interest to test whether the complexity of the Probability Matching is indeed a problem, i.e. are subjects able to use it accurately?

We compare the behavior of three groups of subjects who are ask to perform the same tasks. We only vary the rule used to elicit their beliefs. The tasks to be performed are both cognitive and perceptual tasks. The perceptual task allows us to go further in the analysis. Indeed, signal detection theory which has proved to be a robust model in psychophysics for modeling perception, permits to make prediction about confidence. Signal detection theory relies on two main hypotheses

¹Kadane and Winkler (19988) argues that this elicitation rule may not permit to disentangle beliefs from utilities if the agents’ wealth is correlated to the event. For the tasks we consider, this problem does not hold.

- the visual system encodes the value of perceived visual stimuli with some noise,
- the brain being aware of the visual system characteristics treats the signals received by using a Bayesian analysis.

Therefore, we can compare observed confidences with predicted ones. Beside comparing rules on the basis of how well subjects discriminate, we will also examine how close observed confidence are to predicted confidence.

Our design rules out any strategic interaction, i.e. subjects are paid according to their own performance only. Using simple decision tasks, rather than games, allows us to rule out some possible confounds that arise with strategic interaction (e.g. subjects may coordinate on some strategy that allows them to predict their opponent strategy easily, transforming the initial game into a coordination game). Our subjects have thus to assess their own probability of success to simple tasks that only involve yes/no answers.

Our main finding is that the elicitation rule used does matter. According to various criteria, the Probability Matching rule provides more accurate beliefs. As this rule also has the advantage of not being sensitive to risk aversion, it has a clear advantage over the Quadratic Scoring Rule. The Free rule does a decent job at eliciting beliefs and even outperform the Quadratic Scoring Rule according to various criteria. These results suggest that eliciting rules introduce more or less discrepancy in the translation of evidence signal used to perform task into stated confidence. A possibility we envisaged is the fact that subjects may learn how to use the rule. The design of experiment permits to test the presence of such learning effects. We find some mild evidence of a progress in discrimination for the Probability Matching rule and the Quadratic Scoring Rule but it does not modify our conclusion.

Above the particular problem of elicitation rule, discrepancy in the translation of evidence signal into stated confidence may be linked to some intrinsic ability to form accurate confidence feeling from evidence signal. Since we use two very different tasks, we can study whether we find some pieces of evidence indicating some characteristic properties of judgment to form accurate confidence. It is indeed the case.

In what follows, we briefly recall their main properties and describe a simple experimental protocol aimed at comparing them. In section 4 we introduce methods for comparing the elicitation rules. In section 5, we used the data to run the comparison. In section 6, compare confidence accuracy between tasks. Section 7 concludes.

2 Elicitation Rules

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

2.1 Proper Scoring Rules

Experimental economics proposes some methods to measure the subject's beliefs. Following the standard approach of paying subjects according to their decision, the belief's elicitation is incentivized in a way such that the payoff of a subject increases with the agreement between stated beliefs and realized events. A scoring rule is defined as *proper* if it is incentive compatible, i.e. if subjects can't obtain a higher expected payoff by reporting a probability different than

the true one. In the literature, we can find some proper scoring rules as the quadratic (Brier, 1950), the logarithmic (Good, 1952), or the spherical one (Winkler and Murphy, 1968).

In experimental economics, the most commonly used proper scoring rule is the quadratic rule². In its most simple version, when only two outcomes "success" or "failure" are considered, the Quadratic Scoring Rule consists of 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 nature³. This extends to more general cases where there are n possible outcomes

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

where $I_{i,k}$ takes value 1 if $i = k$ and 0 elsewhere. Remark that this theoretical presentation which is standard refers heavily to some subjective probabilities a subject has in mind. As noted in the introduction, this explicit reference is not necessary in the instructions provided to the subject.

Note also that this rule guarantees a sure paiement when subjects report the same probability for each possible outcome. Thus (very) risk averse subjects may prefer a sure paiement rather than a risky one, and will not report their beliefs truthfully to limit the risk. This is the major problem with the quadratic scoring rule, since it is an incentive compatible scoring rule only under the assumption that subjects are risk neutral. But this assumption of risk neutrality is inconsistent with many experimental results, as most experimental subjects appear to be risk averse. This is a well known weakness of the Quadratic Scoring Rule, but, to the best of our knowledge, there are no assessment of the importance of this limit. (*Voir siii on laisse cette phrase ou s'assurer que Offerman ne le fait pas*).

2.2 The Probability Matching rule

In this experimental study, in order to elicit the level of confidence, we use a procedure (henceforth, the *Probability Matching rule*) put forward by Holt Holt (2006) Holt and Smith (2009) and inspired by an experiment of Grether Grether (1992). 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 she succeeds and a smaller reward $F < S$, if she fails. If the true probability of success is p , the subject should be willing to exchange her task lottery for any lottery that provides a reward of S with probability $q > p$ (and reward F with probability $1 - q$). So now consider the following mechanism: after the subject reported a probability p , a random number q is drawn. If q is smaller than p , the subject is payed according to the task lottery. If q is greater than p , the subjects is payed according to a new lottery that provides the same reward with probability q , called the bonus lottery.

The Probability Matching rule is very much in the spirit of the Becker-DeGroot-Marshak (1964) mechanism and does provide incentives to truthfully reveal p . To make this clear, suppose that the subject thinks the true 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 subjects

²Nyarko and Schotter Nyarko and Schotter (2002), Offerman, Sonnemans, Van de Kuilen, and Wakker (2009), or Palfrey and Wang (2009))

³Note that under this rule, a subject who reports a value of .5 for both probability will get a sure score of .5. Suppose now that he reports a probability of success of .7, he thus gets a gamble with .7 chance to get .91 and .3 chance to get 0.51. This results in an expected gain of .79.

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. So underreporting p makes the subject worse off with positive probability. Now consider a subject who overreports by stating a $r > p$. Again, the interesting case arises when $r > q > p$. In such 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. All in all, the subject has then incentives to truthfully report his best estimates.

The true advantage of the Probability Matching rule is that even risk averse or risk seeking subjects receive incentives to truthfully report the estimated probability. More details and a formalization can be found in Karni (2009).

2.3 The "Free" rule

The Free rule simply consists in asking subjects to report their beliefs, without attaching any monetary consequences to stated probabilities. Nothing is done to provide incentives. Note that there are no incentives neither to misreport as the consequences are the same as reporting the "true" probability. The strong advantage of such a rule is of course its simplicity. It is the less cognitively demanding one, especially compare 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 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.

3 Experimental Design and Procedures

In this section we present the design and provide details about the tasks involved.

3.1 Experimental design

The experiment took place at Laboratory of Experimental Economics in Paris (LEEP). Subjects were recruiting using LEEP's database. Most subjects are students in all area. The experiments last for about 90 minutes. Subjects were paid 19 € on average. The experiment was computer based using Matlab with the Psychophysics Toolbox version 3 (Brainard Brainard (1997)) and the experiments have been done on computers with 1024x768 screens.

Our experimental design is based on two tasks that are described in the next subsection. The important feature of our protocol is to use three blocks of tasks. The first block is composed of quiz questions. Beliefs are elicited but no feedback is provided. Then subjects move on to the perceptual task in which they get direct feedback. This should help subjects improving their use of the elicitation rule. The third block is composed of quiz questions, 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 are comparable (similar subjects and similar success rates)

3.2 Tasks

Our experiment involves two tasks which includes beliefs elicitation, a cognitive one and a perceptive one. In addition, we included some preliminary questions to measure some psychometric variables and risk aversion.

The cognitive task is composed of general knowledge questions, i.e. quiz questions, with "yes" or "no" answers. Typically the questions are similar to the following ones: "*Is the distance between London and Tokyo greater than 12 000 km?*". In this experiment, we use two blocks of questions including 36 questions each. So, each subjects has to answer a total of 72 quiz questions. A special attention was devoted so that each block is of similar difficulty.

The perceptual part is unusual in experimental economics but often used in psychophysics. The aim of this task is to compare the number of dots contained in each circle (see figure 1). The two circles are only display for a short period of time, about 1s, so it is impossible to count the dots. Five levels of difficulty are used, 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)). Compare to the quiz questions, this offers a way of controlling the difficulty of the task according to individual skills. Furthermore, as perceptual tasks are fast, it allows us to get a high number of trials. In this experiment, subjects performed some preliminary trials, used to calibrate the difficulty of the task, and then run 100 trials.

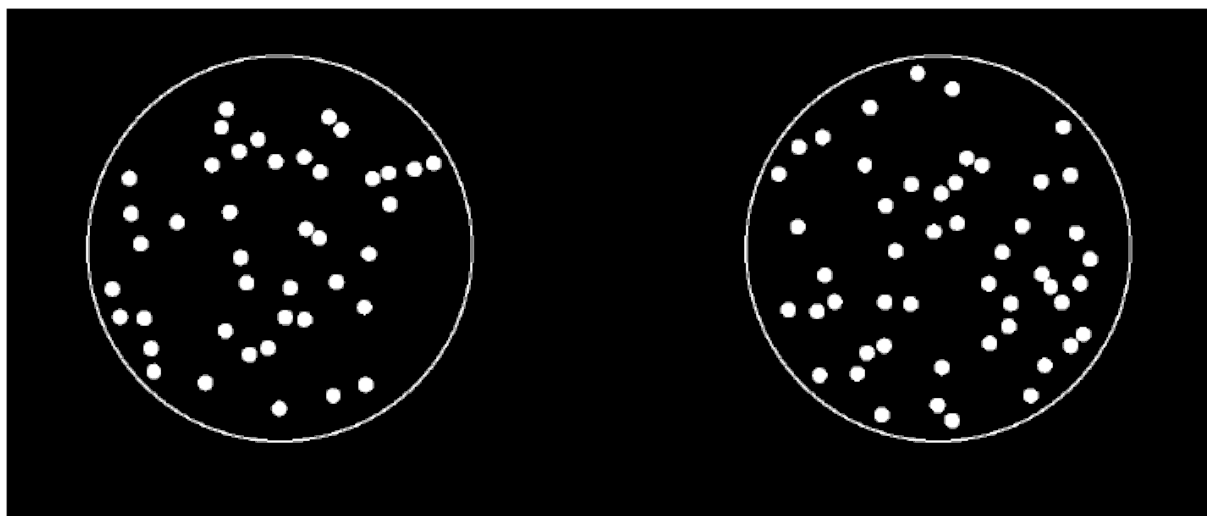


Figure 1: Perceptual task

3.3 Elicitation Procedures and payments

The general principles of each elicitation rule are described in the previous section. Here we only provide details about the way each rule was implemented in our experiment. We also provide details about the payments.

The cognitive tasks are paid according to a standard procedure to avoid edging problems: one question is selected at random at the end of the experiment and payments are computed

according to elicitation rule used. The case of the perceptual task is different, as each successful trial is reward 10cts. Subjects also received a show-up fee of 5 €.

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

Payments under the quadratic scoring rule are given in table 1. Subjects can thus get a sure paiement of 7.5€ or take greater risks, e.g. receive 10€ if their choice is correct, but 0€ if they failed. Note that the corresponding probability were not reported. Indeed, we did not tell them 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 preference revealed approach and reduces confusion with the free rule.

The probability matching rule is implemented using a 0 to 100 scale, with steps of 5. Payments are determined according to the procedure previously described. 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 has no 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 if they win or not. Remark that in such case, their payment does not depend any longer on the quality of their answer. A favorable draw also lead to a paiement of 10€.

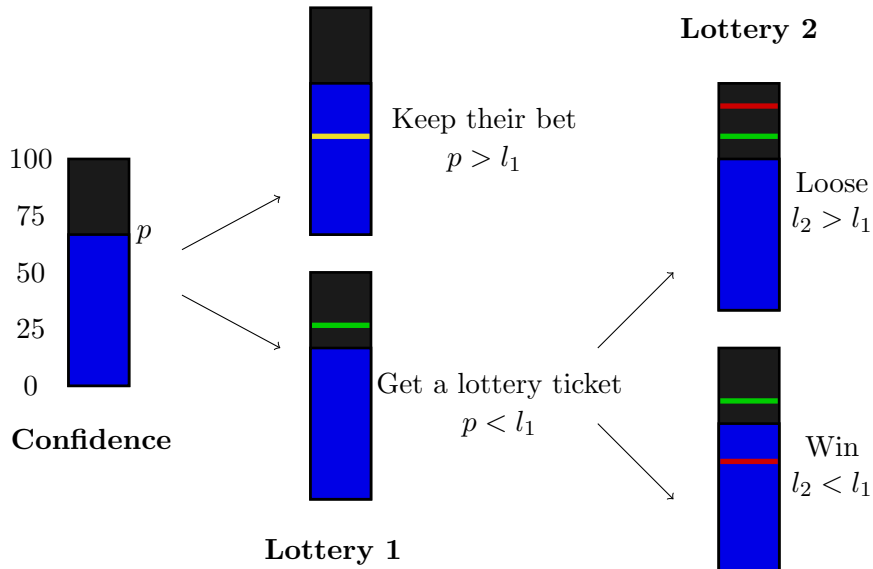


Figure 2: Probability Matching

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

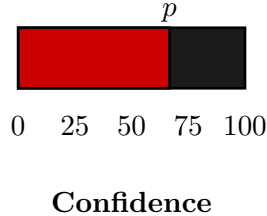


Figure 3: Free Rule

Note that despite we use tasks with binary choices, the scale of confidence is from 0 to 100. One can argue that a half scale from 50 to 100 is more relevant because if the confidence in an alternative is less than 50% subjects should switch to the other choice. Subjects were instructed that a 50% answer provides the highest expected payoff when they have no cue on what the correct answer is. However, reporting values below 50% could be justified, in particular if (1) subjects are sensitive to ambiguity aversion or (2) are sensitive to a competence effect, since we learned from Heath and Tversky (1991) that subjects prefer to be paid on the basis of their skill when confident (i.e. on the basis of the task lottery), and prefer a random paiement (i.e. the bonus lottery) when they are not. According to Heath and Tversky, this phenomena supports the view that success can be attributed to competence and failure to bad luck. Although we don't expect a strong effect in that direction, some subjects may report subjective probability below 50 for these reasons.

4 Methods for comparing elicitation rules

In this section, we first present succinctly how we apply signal detection theory and then we present three criteria based on statistical tools that are useful to compare elicitation rules.

4.1 Signal detection model

The starting point for signal detection theory is that nearly all reasoning and decision making takes place in the presence of some uncertainty. The basic idea is that the subject receive a noisy signal provided by the sensory system which treats the stimuli. In the perceptive task, the subjects have to compare the number of dots contained in two circles. Since the display of circle with dots last just for a short period time, the vision system sends a noisy signal to the brain. A basic model assume that if x is the number of dots, the vision system sends a quantitative signal y which follows a normal law $\mathcal{N}(x, \sigma_i^2)$ where σ_i reflects the sensibility quality of agent i vision system. Therefore, when observing two circles with respectively x_L and x_R dots (L and R stands for left and right), the brain receives and compares two signals y_L and y_R . He guesses left in case $y_L > y_R$. Given the real difference $\tilde{x} = x_L - x_R$, the brain receive a $\tilde{y} = y_L - y_R$

difference in signal which follows a normal law $\mathcal{N}(x_L - x_R, \sigma_i^2)$. In the task, there was always a circle which contains 50 dots and the second circle contains $50 \pm \alpha_j$. Five levels of difficulty were defined. Two were the same for all subjects: 1) level $\alpha_0 = 0$: the two circle contains 50 dots and success was drawn at random, 2) level $\alpha_4 = 25$: the second circle contains 75 dots. Three levels were intermediary and adapted to each subject. During the training part, the medium difficulty level α_2 was adjusted to a value in order to make the subject succeed in 70% of the case at this level. That means that α_2 is such $F(0|\alpha_2, \sigma_i^2) = .3$ where F is the cumulative distribution for the normal law. Table 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 α_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 α_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 α_2 . Then if the model predicts that we should observe a 59 % success rates at level α_1 and 82 % at level α_3 . Compared to the observed success rates which stand at 59% and 80% respectively, we see that the model is quite robust.

Signal detection theory assumed that the brain system is able to run bayesian analysis of perceptive signals. We use this approach to modelize confidence in order to predict expected distribution of confidence. The idea is the following (see the appendix for more detail). Given the signal received \tilde{y} , by using Bayes law the agent evaluate his probability of being right. To apply Bayes law, we make the assumption that the agent is aware of the probability distribution of dots used during the task and of his own sensibility quality σ_i . Then, 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 criterias presented hereafter.

4.2 Calibration

At the empirical level, elicitation rules can be compared according to several criteria. The most commonly used criteria is the distance between the mean predicted success rate and the actual one. This is the so called calibration criteria. Well calibrated subjects are those who exhibit a small distance between means that subjects, on average, accurately predict their success rate. As the beliefs we are interested in are relative to self-confidence, we expect to find some overconfidence in our data. It is thus likely that individuals predict higher success rate that the one they really obtain. We will discuss this point more in depth in the corresponding section.

The measure of calibration is relatively straightforward. Consider a subject who stated beliefs about n events, p_i , being his stated probability for event i , x_i being a variable that take value 1 is he accurately predicts event i .

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

A calibration of 0 indicates a subject who is perfectly calibrated. A positive calibration indicates that the subject is overconfident, while a negative one indicates underconfidence.

Note that by construction, the confidence predicted by the signal detection model exhibits perfect calibration: the mean success rate for confidence \tilde{p} is exactly \tilde{p} .

4.3 Discrimination

Another important criteria 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 has very low discrimination ability will provide the same probability whatever the events. It is important to note that such a subject might however be well calibrated if he reports for each trial a probability equal 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 for visualizing, organizing and selecting classifiers based on their performance. 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 one to compute the true positive rate (TPR), i.e. the fraction of predicted success that are correctly predicted, and the false positive rate (FTP), i.e. the fraction of failure 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. The ROC Area is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. It can be shown that the area under the ROC curve is closely related to the Mann–Whitney U (Hanley and McNeil (1982)), which tests whether positives are ranked higher than negatives. It is also equivalent to the Wilcoxon test of ranks. The ROC Area is related to the Gini coefficient (G1) by the following formula: $G1 + 1 = 2ROC\ Area$.

One advantage of the ROC analysis is that it uses confidence level only ordinally. For instance, if a subject is good to rank his confidences but has some problem to give absolute values, his ROC Area can still be high. Furthermore, the ROC Area is immune to distorsion bias induced by the elicitation rules. For instance, the ROC Area for QSR is not plagued by risk aversion.

4.4 Composite index

One important measure of overall performance in the accuracy of a judgment is the *Brier Score* (Brier, 1950). Let p_i be the subjective probability of a subject on an event E_i (e.g. confidence) and x_i be an indicator function that equals 1 if the event E_i occurs and 0 otherwise. The Brier Score is given by

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

where n is the number of elements in P the set of probability assessments. This index gives an overall performance of a judgment with lower values indicate better performance. In order to obtain an identification of the calibration and discrimination concepts we can use the *Murphy*

Decomposition (Murphy, 1972 ; Yates, 1982) of the Brier Score which is expressed as following:

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

where N_p is the number of times that the probability judgment of the designated event equals the confidence category p , f_p is the relative frequency of occurrence in that class, and f is the overall relative frequency of the designated event. This decomposition could also be writing as

$$S = V - DI + CI$$

where V is the variance of the outcome variable which is independent of the judgment, CI the calibration index measuring the difference between the observed hit rate (f_p) and the identity line, and DI is the discrimination index measuring the ability of the hit rate around the overall base rate (f). This index and its decomposition are good measures of the accuracy of confidence assessment. Note that contrary to the ROC analysis, confidence level are used cardinally.

5 Results

To compare elicitation rules, we consider two types of results, those based on observation of confidence, the second based on quality judgment index. First, we pool the data for each rule and examine the distribution of confidence obtained. In the case of the perceptive task, we can compare observed and predicted confidence. Then we will compare elicitation rule according to quality judgment index. Finally, we examine learning effects.

5.1 Confidence

As a preliminary step, we perform some descriptive analysis that allow us to get a general picture of elicited beliefs. In figure 4, we draw the cumulative probability distributions of elicited confidence for both task and for each rule. The results for all subjects are pooled.

It is clear that while the Probability Matching 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 Quadratic Scoring Rule has a strong tendency to have stated probabilities concentrated on two values, 50% and 100%: almost two third of elicited probabilities are among these tow values, this is twice as much as for the two other rules.

For the perception task, we draw in figure 5 the cumulative probability distributions for each rule and for the Signal Detection Model prediction (all subject pooled). The shape of the Probability Matching curve and of the Free curve follows the SDT one while QSR differs a lot.

The difference between the QSR curve and the three others is also due to the concentration of 50% and 100% probabilities with again almost two third of elicited probabilities taking these values. Note that SDT predicts that only 24% of stated confidence should take these two values (see table 12 in appendix). This visual feeling is confirmed by computation of the Chi-square distance between observed and predicted confidence distribution for each group of subjects. To fix the 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 put a probability 1 on 100%. Results are given in Table 2.

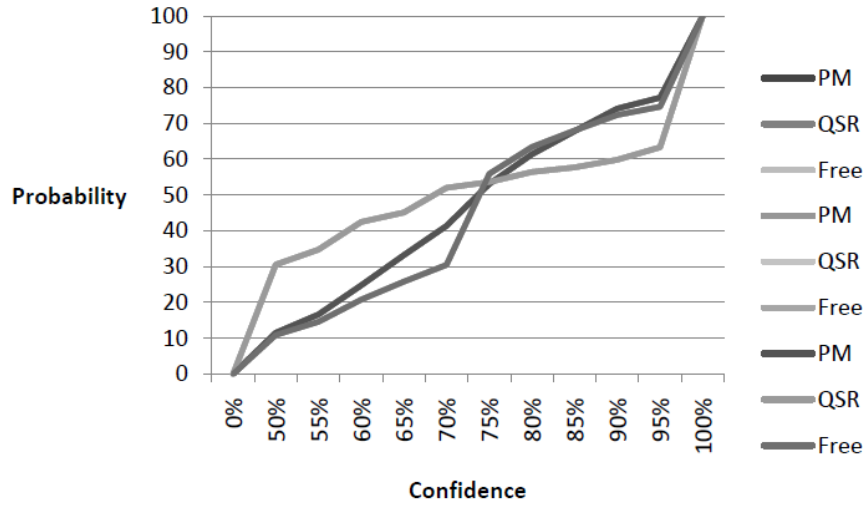


Figure 4: Cumulative probability distribution of confidence for each rule

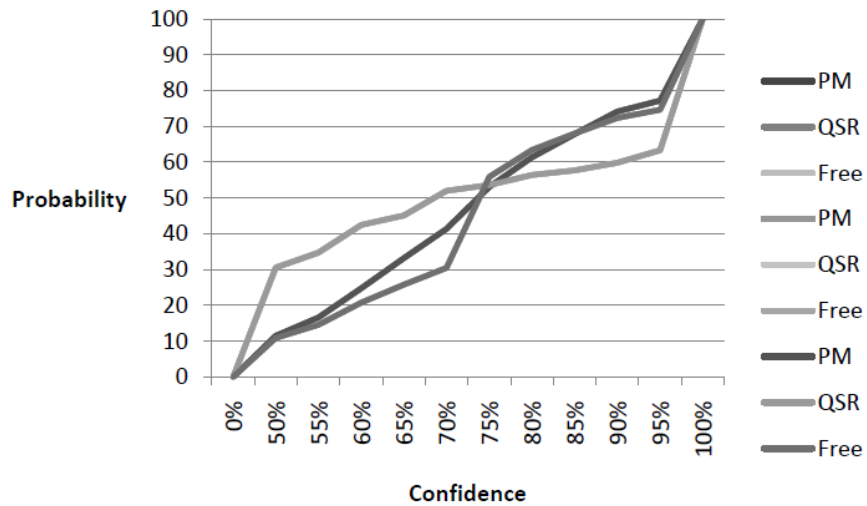


Figure 5: Cumulative probability distribution of confidence for each rule and SDT

Chi-square distance btw. pred. confid. and...	PM (n=38)	QSR (n=35)	Free (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 2: Chi-square distance between confidence distributions

The PM rule clearly outperforms QSR in its ability to fit with predicted confidence.

For the perception task, we make subject follows a training phase during which an automatic adjustment of difficulty was done so as to make the subject succeed at 70% for the medium difficulty level. This adjustment was not perfect and during the main task subjects success rate differ with a mean success of 71.2% overall, a standard deviation of mean success equal to 6% and mean success ranging from 56% to 86%. Accordingly, the Signal Detection Model predict that mean confidence should be lower for those for which the task was more difficult given their ability (see appendix for detail). In the following table, we give the observed and predicted mean confidence value as well as correlation between observed and predicted mean confidence.

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

Table 3: Observed and predicted mean confidence. (Std. in brackets) (* means significance at 10%)

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

5.2 Quality of judgment

Before we compare rules according to synthetic index, we draw a general picture of how subjects, taken together, do judge themselves. Fig.6 compares predicted success rate, according to stated beliefs, to actual success rate. 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 rate gets larger 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 rate are that roughly in line with expected ones. Even if the three rules are similar, some differences worth however noting. None of 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 the one of the QSR, for which there is not significant differences among stated probability in the range [65, 95]. On average, any such probability leads to an approximate rate of success of 67%. As a result we expect poor discrimination in that range of values.

The fact that high confidence intervals lead to strong overconfidence should not be taken at face value. As we used quizz questions, it is the case 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 Fig.7 provides some indication of the frequency of such questions. Removing these misleading questions will thus diminish overconfidence by almost half.

An hard -easy effect is also observed for the perception 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 (α_0 or α_1) because

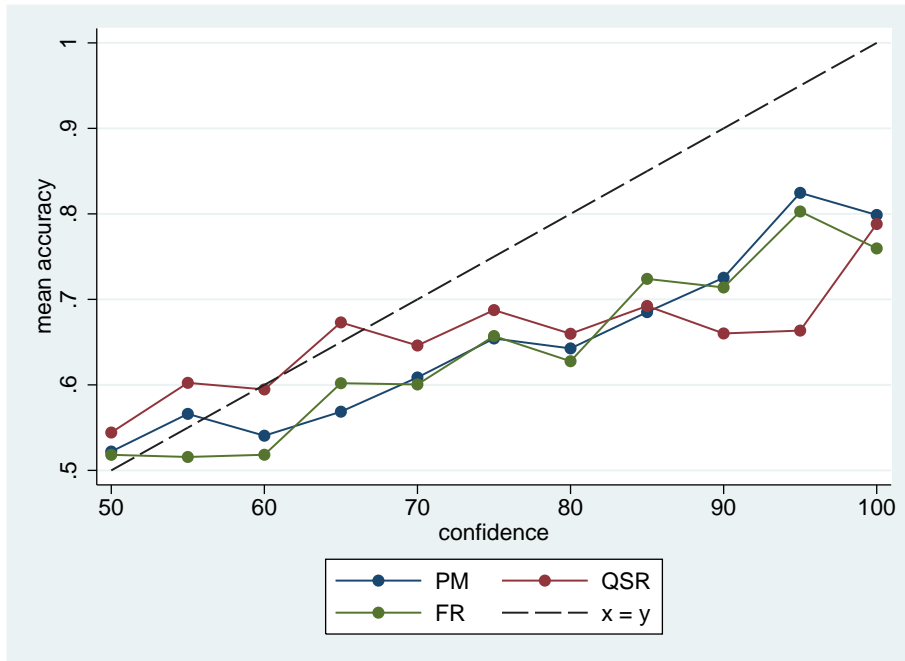


Figure 6: Matching between confidence and accuracy

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 probability matching and 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.

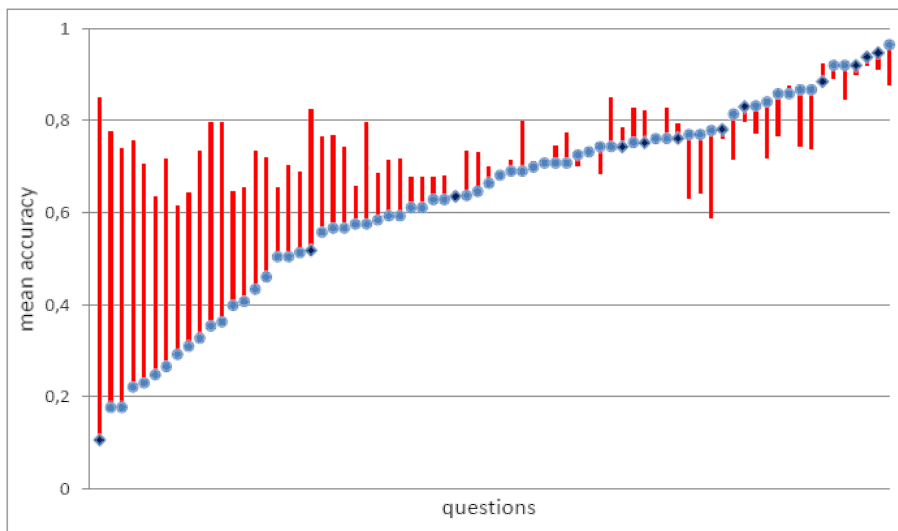


Figure 7: Ranked success rate and overconfidence to quiz questions

This figure shows the level of mean accuracy for each questions 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.

of erroneous signals and underconfident when the difficulty is low (α_3 or α_4). In Table 4, we report observed and predicted confidence as well as the observed success rate.

	Observed mean confidence			Predicted mean confidence			Success rate		
	PM	QSR	Free	PM	QSR	Free	PM	QSR	Free
α_0	71%	69%	72%	64%	63%	64%	45%	52%	50%
α_1	72%	71%	74%	64%	64%	64%	60%	60%	58%
α_2	72%	73%	74%	66%	65%	66%	69%	67%	67%
α_3	77%	76%	78%	71%	69%	70%	82%	79%	80%
α_4	94%	96%	94%	95%	95%	91%	100%	100%	100%

Table 4: Observed and predicted mean confidence. (Std. in brackets)

Remark that the Signal Detection Model predicts very small differences of mean confidence for the three more difficult level while success rate range from 50% (α_0) to 70% (α_2). Empirically, these predictions are confirmed for the three rules.

We turn now to the statistical index quality judgment. Table 5 provides some measure using the statistical index described above. QSR performs better in terms of calibration, displaying a lower degree of overconfidence than the two other rules. As expected, the PM rule provides a better discrimination than the scoring rule with an area under the ROC curve of 0.6401. The Brier score provides an aggregate measure. The PM rule clearly outperform the two other rules with a Brier score of 0.2245. The more surprising results is that the Free rule appears slightly better than QSR.

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

Table 5: 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 rules, we have the value of the Brier score, the level of overconfidence and the area under the ROC curve with the standard deviation. Then, we compare the rules two by two and find and level of difference for each criteria. We perform a test of difference: for the Brier score and the overconfidence we test the significativity of the inequality between the rules (t-test with the p-value in parenthesis); and for the AUC we test the significativity of the equality between the rules (Chi-square test with p-value in parenthesis).

And, as already mentioned, the discrimination is rather poor for the intermediate level of probability. Part of this result can be interpreted as a direct effect of risk aversion. A stated probability of 50% provides a sure paiement, will higher probability means taking a risk. However, this does not explain the poor discrimination associated with higher probability, nor the attractiveness of 100%. An explanation for the underperformance of QSR is the noise that

this rule induces in the process of translating the beliefs used in the performance task in elicited confidence. In the previous section we observed that for the perception task, stated confidence were far from predicted confidence in the case of QSR. It is also the case for discrimination. Indeed, the Signal Detection Model predicts that those for which the task was more difficult given their ability should have a lower discrimination index, either measure by the ROC area or by the Brier score. Furthermore, we expect to observe correlation between observed and predicted discrimination index. Clearly, we observe in table 6 that the level of correlation are poorer for QSR than for the two other rules. For ROC area, there is no correlation for QSR. Among the two other rules, the PM rule is slightly better than the Free rule.

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

Table 6: 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%)

5.3 Learning

Our experimental design offers subjects the opportunity to learn using feedback. Remember that we use three blocks of questions. The second one is perceptual tasks 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 following table provide details about the learning effect under the three different rules. Table 7 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. This is not what we found. The most significant effect, if any, is found for the PM 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 no rule has a clear advantage in term of learning.

6 Inter-task comparison

Up to now, we did not go into the detail of individual data but differences appears either in terms of calibration than in terms of discrimination. Overall, calibration ranges from an underconfidence of -14% to an overconfidence of +39% (Std = 10%) while ROC areas vary from 0.51 to 0.79 (Std. = 0.06) and the Brier scores from 0.14 to 0.39 (Std = 0.04). In economy,

Tasks	Brier Score	Overconfidence	ROC Area
PM_Q1	0.2464	0.1117 (.0131)	0.6158 (.0154)
PM_Q2	0.2548	0.1326 (.0130)	0.6318 (.0151)
PM_(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 7: Learning: cognitive tasks

overconfidence is highly discussed (Camerer, Biaisi) but discrimination is not considered while it is probably an important aspect of behavior. However, we observe some subjects are completely unable to discriminate: their stated confidence are totally uninformative about future success and failure. Formally, calibration and discrimination are statistically independent: one can be very well calibrated without discriminating well and vice versa. One exception is the case of extreme behavior. For instance, a subject who is always 100% sure is overconfident and discriminates badly. One interesting thing to examine is to see whether we find empirical correlations between the two dimensions. Table 8 indicates correlations for subjects whose overconfidence is below 30%.

Corr. btw. Calibration and ROC area	All (104)	PM (38)	QSR (34)	Free (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 8: Calibration versus discrimination

We find significant correlation only for the Quizz task and this is driven by the result obtained using QSR. Since we observe before that QSR seems to perform badly in eliciting confidence, it appears that calibration and discrimination abilities are independently distributed in the population.

Are judgment abilities domain specific? There is some experimental evidence that people are overconfident over domains even if their level of overconfidence vary with the domain, and that more overconfident people in one domain tends to be also more overconfident in other domain (see West and al. (1997)). For discrimination it seems also to be the case that discrimination abilities exports accross domains (see Bornstein and al. (1999)). In table 9, we report our findings on this question.

For all rules, we find some high correlations for calibration. For the QSR, this especially high correlation drives the correlation observed for the Brier scores. For discrimination, we find significant correlation only for PM and for the Free rule. Note that the Signal Detection Model predicts that subjects cannot have a high ROC area if their task was hard in the perception

Corr. btw. Quizz and Perception	All (104)	PM (38)	QSR (34)	Free (32)
Calibration	0.57***	0.47***	0.69***	0.49***
ROC area	0.27***	0.35**	0.15	0.33*
Brier score	0.29***	0.18	0.43**	0.28

Table 9: Correlation between task

task. Thus, to measure the subject intrinsic discrimination ability, we must take into account the predicted ROC area which is a benchmark value. So we create 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 quizz task. In table 10 we report these correlations.

	All (104)	PM (38)	QSR (34)	Free (32)
Corr. btw. quizz ROC area and ROC _{pa}	0.10	0.31*	0.00	0.07

Table 10: Correlation

Significant correlation remains only for PM. This last result indicates two things. First, discrimination ability seems more domain specific than calibration ability. Nevertheless, we see that in order to study such sophisticated question, the choice of elicitation rule matters a lot.

7 Conclusion

The choice of a particular elicitation rule does matter. If we concentrate only on discrimination and calibration index, these differences are not so big but still significant. Signal Detection Theory permits to make finer prediction about beliefs formation and to make forecasts about what we expect to observe. What is striking is that this model which is known to be very robust to predict performance and behaviors generates predictions that fit quite well with the data obtained with the Probability Matching rule and not so badly with the Free rules' data. With QSR, the fit is clearly poorer. In our view, these results show two things: Signal Detection Theory which assumes that beliefs signals are formed when performing task is a good model and that the elicitation rules which ask subject to reports their feelings in terms of a visual metric outperform elicitation rule which are based on a revealed preference approach through the choice of stakes. Incentives is not the main issue even if we can suppose that incentives explain why Probability Matching outperforms teh Free rule. For experimental needs, if one expects some precise values, then QSR does not seem to be a good candidate. QSR is the more popular rule in experimental game theory. A disturbing question arises when we compare the relative performance of subjects in pure decision tasks, like the ones presented here, to the case of strategic interaction where subjects have to form belief about their opponent's behavior. Costa-Gomes and Weizsacker (2008) for example found that subjects perform poorly in predicting the

behavior of another player in three by three games (see appendix B of their paper for more details). In our decision tasks, subjects obtain better Brier scores than someone who does not discriminate at all by stating probability of 0.5 for every question. The same is not true for the subjects who participated in Costa-Gomes and Weizsacker, they are doing, on average, worst than subjects that assign to each possible outcome the same probability. Is it a problem of poor strategic decision thinking or the Quadratic Scoring Rule is part of the problem?

8 Appendix

8.1 Signal detection model

We detail here how we apply the model.

8.1.1 The basis for the prediction of confidence distribution

Consider for instance a subject who receive a signal $\tilde{y} = y_L - y_R > 0$ and thus who guesses left. The signal \tilde{y} follows a normal law with a mean equal to the real difference in dots $\tilde{x} = x_L - x_R$ and a variance σ_i^2 where σ_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.

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 catch the probability of evenness between x_L and x_R in which case the subject win 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 receive a signal \tilde{y} such that $P(|\tilde{y}|) = \tilde{p}$. Given α_j , the density function for the confidence \tilde{p} is equal to

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

In the experiment, confidence was elicited with a path of 5%. We do the same 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 α_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 reflect perfect calibration, that is, the mean success rate when confidence is \tilde{p} is equal to \tilde{p} :

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

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 σ_i . His ability is revealed through his success rate at levels $\alpha_{j=1,\dots,4}$. At level α_j , we observe $n_{i,j}$ trials and $r_{i,j}$ successes. We compute σ_i such that

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

The following table gives the descriptive statistics for the σ_i .

	Nb	Mean	Std. Dev.	Min;Max
σ_i	108	7.95	5.30	2.8 ;50.5

Table 11: Ability

From σ_i , we can compute for each i and level α_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:

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

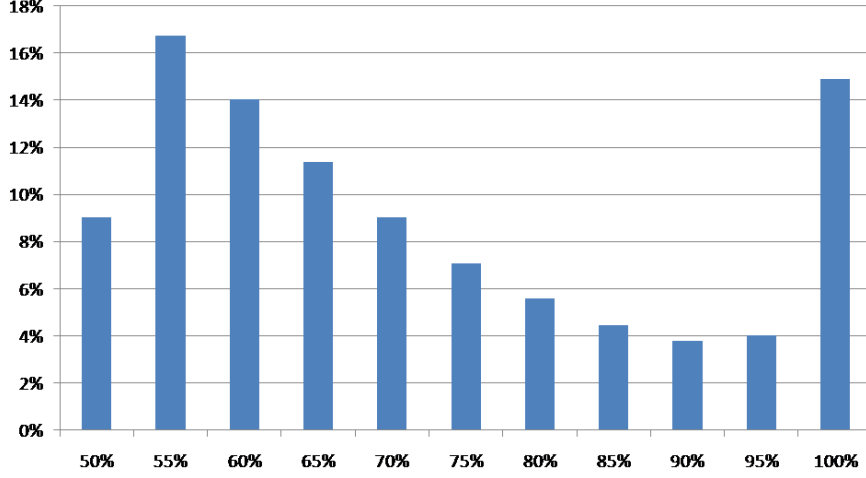


Figure 8: SDT prediction of confidence distribution

Some descriptive statistics for the confidence distribution are given in the following figure.

Given these confidence distributions, we can calculate 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 α_0 may not be exactly 50% because of a low number of trials and from the approximation done in the estimation of confidence. We can check in the next table that it is indeed the case.

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 by subject.

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

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

The following table 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

- ANDERSEN, S., J. FOUNTAIN, G. HARRISON, AND E. RUTSTROM (2007): “Eliciting Beliefs: Theory and Experiments,” *Working Paper*.
- BARANSKI, J., AND W. PETRUSIC (1994): “The Calibration and resolution of confidence in perceptual judgments,” *Perception and Psychophysics*, 55(4), 412–428.
- (1999): “Realism of confidence in sensory discrimination,” *Perception and Psychophysics*, 61(7), 1369–1383.
- BENOIT, J., AND J. DUBRA (2008): “Overconfidence?,” *MPRA Paper No. 6017*.
- BENOIT, J., J. DUBRA, AND D. MOORE (2009): “Does the Better-Than-Average Effect Show That People Are Overconfident?: An Experiment,” *MPRA Paper No. 13168*.
- BHATT, M., AND C. CAMERER (2005): “Self-Referential Thinking and Equilibrium as States of Mind in Games: fmri Evidence,” *Games and Economic Behavior*, 52(2), 424–459.
- 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.
- BJORKMAN, M., P. JUSLIN, AND A. WINMAN (1993): “Realism of confidence in sensory discrimination: The underconfidence phenomenon,” *Perception and Psychophysics*, 54(1), 75–81.
- BLANCO, M., D. ENGELMANN, A. KOCH, AND H.-T. NORMAN (2008): “Belief elicitation in experiments: Is there a heging problem?,” *IZA Discussion paper No. 3517*.
- BLAVATSKYY, P. (2009): “Betting on own knowledge: Experimental test of overconfidence,” *Journal of Risk and Uncertainty*, 38(1), 39–49.
- BRAINARD, D. (1997): “The Psychophysics Toolbox,” *Spatial Vision*, 10, 433–436.
- BURKS, S., J. CARPENTER, G. L., AND A. RUSTICHINI (2009): “Is Overconfidence a Judgment Bias? Theory and Evidence,” *Working Paper*.

- CAMERER, C., AND R. HOGARTH (1999): “The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework,” *Journal of Risk and Uncertainty*, 19(1-3), 7–42.
- CESARINI, D., O. SANDEWALL, AND M. JOHANNESSON (2006): “Confidence interval estimation tasks and the economics of overconfidence,” *Journal of Economic Behavior and Organization*, 61(3), 453–470.
- CLARK, J., AND L. FRIESEN (2009): “Overconfidence in forecasts of Own Performance: An Experimental Study,” *The Economic Journal*, 119(534), 229–251.
- 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.
- EREV, I., T. WALLSTEN, AND D. BUDESCU (1994): “Simultaneous over and underconfidence: The role of error in judgment processes,” *Psychological review*, 101(3), 519–527.
- FAWCETT, T. (2006): “An introduction to ROC analysis,” *Pattern Recognition Letters*, 27, 861–874.
- FIGORE, A. (2009): “Experimental Economics: Some Methodological Notes,” *MPRA Paper No. 12498*.
- FISCHHOFF, B., P. SLOVIC, AND S. LICHTENSTEIN (1977): “Knowing with Certainty: The Appropriateness of Extreme Confidence,” *Journal of Experimental Psychology: Human Perception and Performance*, 3(4), 552–564.
- FRIEDMAN, D., AND S. SUNDER (1994): *Experimental Methods - A Primer for Economists*. Cambridge University Press.
- GIGERENZER, G., U. HOFFRAGE, AND H. KLEINBOLTING (1991): “Probabilistic Mental Models: A Brunswikian Theory of Confidence,” *Psychological Review*, 98(4), 506–528.
- GNEITING, T., AND A. RAFTERY (2007): “Strictly proper scoring rules, prediction, and estimation,” *Journal of the American Statistical Association*, 102(477), 359–378.
- GOODIE, A. (2003): “The Effects of Control on Betting: Paradoxical Betting on times of high confidence with low value,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(4), 598–610.
- GREEN, D. M., AND J. A. SWETS (1966): *Signal detection theory and psychophysics*. John Wiley and Sons.
- GREYER, D. (1992): “Testing Bayes rule and the representativeness heuristic: Some experimental evidence,” *Journal of Economic Behavior and Organization*, 17, 31–57.
- GRIFFIN, D., AND A. TVERSKY (1992): “The weighing of evidence and the determinants of confidence,” *Cognitive Psychology*, 24(3), 411–435.
- GUALA, F. (2005): *The Methodology of Experimental Economics*. Cambridge University Press.

- 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*.
- HEATH, C., AND A. TVERSKY (1991): “Preference and Belief: Ambiguity and competence in choice under uncertainty,” *Journal of Risk and Uncertainty*, 4(1), 5–28.
- HERTWIG, R., AND A. ORTMANN (2001): “Experimental practices in economics: A methodological challenge for psychologists?,” *Behavioral and Brain Sciences*, 24, 383–451.
- HOELZL, E., AND A. RUSTICHINI (2005): “Overconfident: Do you put your money on it?,” *Economic Journal*, 115(503), 305–318.
- 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*.
- HURLEY, T., AND J. SHOGREN (2005): “An experimental comparison of induced and elicited beliefs,” *Journal of Risk and Uncertainty*, 30(2), 169–188.
- KARNI, E. (2009): “A Mechanism for eliciting Probabilities,” *Econometrica*, 77(2), 603–606.
- KEREN, G. (1988): “On the ability of monitoring non-veridical perceptions and uncertain knowledge: some calibration studies,” *Acta Psychologica*, 67(2), 95–119.
- (1991): “Calibration and probability judgments: conceptual and methodological issues,” *Acta Psychologica*, 77(3), 217–273.
- (1999): “On The Calibration of Probability Judgments: Some Critical Comments and Alternative Perspectives,” *Journal of Behavioral Decision Making*, 10(3), 269–278.
- KLAYMAN, J., J. SOLL, AND S. GONZÁLEZ-VALLEJO, C. AND. BARLAS (1999): “Overconfidence: it depends on How, What, and Whom you ask,” *Organizational Behavior and Human Decision Processes*, 79(3), 216–247.
- KORIAT, A., S. LICHTENSTEIN, AND B. FISCHHOFF (1980): “Reasons for Confidence,” *Journal of Experimental Psychology: Human Learning and Memory*, 6(2), 107–118.
- KORIAT, A., L. SHEFFER, AND H. MA’AYANT (2002): “Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice,” *Journal of Experimental Psychology: General*, 131(2), 147–162.
- LAHIRI, K., AND G. WANG (2007): “Evaluating probability forecasts for GDP declines,” *Working paper, University of Albany - SUNY*.

- LEVITT, H. (1971): “Transformed up-down methods in psychoacoustics,” *Journal of the Acoustical Society of America*, 49, 467–477.
- LIBERMAN, V., AND A. TVERSKY (1993): “On the evaluation of probability judgments: Calibration, resolution, and monotonicity,” *Psychological Bulletins*, 114, 162–173.
- 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 P. L. (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.
- MERKLE, E. (2009): “The disutility of the hard easy effect in choice confidence,” *Psychonomic Bulletin and Review*, 16(1), 204–213.
- MERKLE, E., AND T. VAN ZANDT (2006): “An application of the Poisson race model to confidence calibration,” *Journal of Experimental Psychology: General*, 135(3), 391–408.
- 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* (forthcoming).
- PAESE, P., AND J. SNIJEK (1991): “Influences on the appropriateness of confidence in judgment: Practice, effort, information, and decision-making,” *Organizational Behavior and Human Decision Processes*, 48, 100–130.
- PALFREY, T., AND S. WANG (2009): “On Eliciting Beliefs in Strategic Games,” *Journal of Economic Behavior and Organization*, 71(2), 98–109.
- PETRUSIC, J., AND W. BARANSKI (1997): “Context, Feedback, and the Calibration and Resolution of Confidence in Perceptual Judgments,” *The American Journal of Psychology*, 110(4), 543–572.
- PULFORD, B. (1996): “Overconfidence in Human Judgment,” Ph.D. thesis, University of Leicester.
- PULFORD, B., AND A. COLMAN (1997): “Overconfidence: Feedback and item difficulty effects,” *Personality and Individual Differences*, 23(1), 125–133.
- READ, D. (2005): “Monetary incentives, what are they good for?,” *Journal of Economic Methodology*, 12(2), 265–276.
- RUTSTROM, E., AND N. WILCOX (2009): “Stated versus inferred beliefs: A methodological inquiry and experimental test,” *Games and Economic Behavior* (in press).
- SELTEN, R. (1998): “Axiomatic characterization of the quadratic scoring rule,” *Experimental Economics*, 1, 43–62.

- SNIEZEK, J., P. PAESE, AND F. SWITZER III (1990): “The effect of choosing on confidence in choice,” *Organizational Behavior and Human Decision Processes*, 46, 264–282.
- STONE, E., AND R. OPEL (2000): “Training to Improve Calibration and Discrimination: The Effects of Performance and Environmental Feedback,” *Organizational Behavior and Human Decision Processes*, 83(2), 282–309.
- SUANTAK, L., F. BOLGER, AND W. FERRELL (1996): “The Hard Easy Effect in Subjective Probability Calibration,” *Organizational Behavior and Human Decision Processes*, 67(2), 201–222.
- TANNER, J., P. WILSON, AND J. A. SWETS (1954): “A decision-making theory of visual detection,” *Psychological Review*, 61(6), 401–409.
- TSAI, C., J. KLAYMAN, AND R. HASTIE (2008): “Effects of amount of information on judgment accuracy and confidence,” *Organizational Behavior and Human Decision Processes*, 107, 97–105.
- YANIV, I., F. YATES, AND K. SMITH (1991): “Measures of Discrimination Skill in Probabilistic Judgment,” *Psychological Bulletin*, 110(3), 611–617.