Confidence Measurement in the Light of Signal Detection Theory

Thibault Gajdos[1,2], Sébastien Massoni[3,4,5], and Jean-Christophe Vergnaud[2,3]

[1]*GREQAM, Aix Marseille University, EHESS*

[2]*CNRS*

[3]*Centre d'Economie de la Sorbonne, University of Paris 1*

[4]*Paris School of Economics*

[5]*QuBE, Queensland University of Technology*

ABSTRACT

We compare three alternative methods for eliciting retrospective confidence in the context of a simple perceptive task: the Simple Confidence Rating (a direct report on a numerical scale), the Quadratic Scoring Rule (a post-wagering procedure) and the Matching Probability (a generalization of the no-loss gambling method). We systematically compare the results obtained with these three rules to the theoretical confidence levels that can be inferred from performance in the perceptive task using Signal Detection Theory. We find that the Matching Probability provides better results in that respect. We conclude that Matching Probability is particularly well suited for studies of confidence that use Signal Detection Theory as a theoretical framework.

Keywords: Confidence, Scoring Rules, Psychophysics, Signal Detection Theory, Incentives, Methodology.

1. Introduction

Humans and animals are able to retrospectively evaluate whether they have made or not the right decision (e.g., in perceptive, learning or memory tasks). This metacognitive ability plays an important role in learning and planning future decisions (Dunlosky & Metcalfe, 2008). For instance, humans are not only able to decide whether a visual stimulus did appear or not, but also to say how confident they are in their answer. Such retrospective judgements are often labelled “type 2 tasks”, as opposed to “type 1 tasks” which consist in discriminating between perceptual stimuli.

In the last years, considerable progresses had been made in the understanding of behavioural (Smith et al., 2003) and neuronal (Fleming & Dolan, 2012) properties of retrospective confidence. These progresses rely to a large extend on a Bayesian analysis of confidence formation (Kepecs & Mainen, 2012). Since Green and Sweets (1966)’s classical book, Signal Detection Theory (SDT) has been routinely and successfully used in experimental psychology to study simple perceptual decisions. It is postulated that stimuli are perceived as a noisy signal by the sensory system. The type 1 task thus reduces to deciding, on the basis of the observation of a random signal (on some internal axis), whether this observation is due to noise or to the presence of the stimuli. This reasoning can be pushed one step further to infer confidence, defined as the probability of having made the correct decision (Galvin et al., 2003). Let us call this model the SD-Confidence model (where "SD" stands for "Signal Detection"). A crucial feature of the SD-Confidence model is that it allows predictions of retrospective confidence levels based on the observation of type 1 decisions. Individuals’ ability to make retrospective confidence judgements can thus be measured by comparing their actual confidence to that predicted by the SD-Confidence model (Galvin et al., 2003; Maniscalco & Lau, 2012). In other words, SD-Confidence model provides a framework in which one can measure inter-individual (Fleming et al., 2010) and inter-tasks (McCurdy et al., 2013; Song et al., 2011) variations in the quality of confidence judgements. In that respect, SD-Confidence model played a crucial role in recent advances in the study of retrospective confidence.

It is not entirely obvious, however, how confidence should be measured. Roughly speaking, three methods are available. The first method consists in asking individuals to explicitly report their confidence, either on a verbal or numerical scale (Dienes, 2007). Such a method, known as Simple Confidence Rating is straightforward and easy to use. Yet, it cannot be used

for non-humans and children. Furthermore it had been argued that it could lead to biased reports, as individuals have no incentives to reveal their "true" confidence, which might require an effort. It has thus been proposed to use an indirect method, called post-decision wagering, based on individual willingness to bet on the quality of their answers (Persaud et al., 2007). For instance, after having made her type 1 decision, the subject is asked whether she is ready to bet €10, €20 or €50 on being right. The idea is that subjects will choose higher stakes when their confidence is higher. Such a method provides incentives, and can be used for non-humans (Middlebrooks & Sommer, 2011). However, it is also affected by individuals' attitude towards risk (Dienes & Seth, 2010). Therefore, using post-decision wagering could lead to erroneous measures, insofar as variations in confidence measured by post-wagering method could to some extend reveal heterogeneity in attitude towards risk, and not in confidence. Dienes and Seth (2010) proposed a method, called "no-loss gambling", that provides incentive and is immune to risk aversion. It consists in asking individuals whether they prefer to be paid according to the correctness of their answer of according to a specified lottery. For instance, individuals are asked whether they prefer to receive €10 in case of success and get nothing otherwise, or to toss a coin and receive €10 if it turns Heads and nothing otherwise. The idea here is that if a subject prefers to be paid according to the correctness of her answer, it indicates that her confidence in her decision is higher than 50%.

These methods have been extensively discussed in the context of the measurement of consciousness, although the question whether confidence ratings are appropriate measures of consciousness is hotly debated (Sandberg et al., 2010). The reasoning goes as follows. Consider a person who performs a simple two-alternative perceptive task, and whose success rate is above the chance level. This means that she at least partially observed the stimulus and used it to make her decision. However, it might be that she is not aware of having perceived the stimulus, and that the whole process is totally unconscious. In this case, she will presumably not be able to discriminate between successful and erroneous trials. Therefore, her retrospective confidence will not be correlated to her actual success. On the other hand, if she is conscious of having used some information, she will be able to at least partially discriminate between success and failures. In this case, her confidence will be correlated with success. Moreover, it is postulated that high confidence levels reveal conscious access to some information. Therefore, type 2 discrimination sensitivity (known as *d'*) will be used as a measure of consciousness. The first important thing we can learn from these studies is that different methods for measuring confidence yield different results (Overgaard & Sandberg,

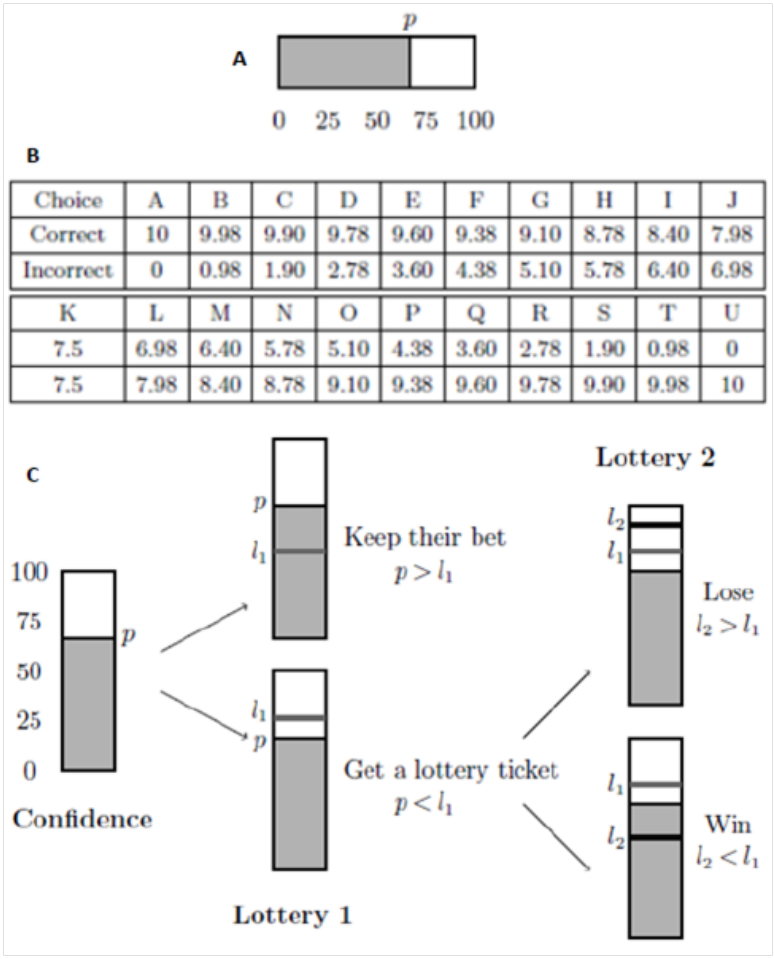
2012). Thus, the choice of the measure matters. On what criterion should one decide that a measure is better than another? Notice that in these studies, the aim is not to measure confidence *per se*, but confidence as a measure of consciousness. Thus, one can reasonably argue that the most sensitive the measure, the better it is.

In contrast, it is difficult to see why type 2 *d'* should be a good criterion for evaluating elicitation rules if one is interested in confidence *per se*, Actually, it seems pretty clear that the appropriate criterion should depend on what we mean by "confidence". In other words, one could not propose a criterion without providing a precise definition of confidence. This is exactly the role the SD-Confidence model plays in studies of metacognition. We should thus rephrase our question: What is the best way to measure confidence, if we want to measure the sort of confidence described by the SD-Confidence model? A possible answer, based on the ideal observer paradigm (Geisler, 2011) could be to retain the measure that provides a confidence as close as possible to that predicted by the SD-Confidence model. This is precisely the approach we follow. The aim of this study is to compare generalizations of Simple Confidence Rating, Post-Decision Wagering and No-Loss Gambling in the light of the predictions of the SD-Confidence model.

2. Method

2.1. Elicitation rules

The main objective of our experiment is to compare three elicitation rules (see Figure 1): the Simple Confidence Rating (SCR), the Quadratic Scoring Rule (QSR) and the Matching Probability (MP). The SCR is a direct report on a numerical scale. The QSR is a fine-grained version of the post-wagering procedure. The MP is a multi-level extension of the no-loss gambling method proposed by Dienes and Seth (2010). This section is devoted to the presentation of these rules, discussion of their main theoretical properties, and the presentation of their experimental implementation.



Choice	A	B	C	D	E	F	G	H	I	J
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

K	L	M	N	O	P	Q	R	S	T	U
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

Figure 1: Elicitation mechanisms for confidence judgments. (A) represents the Simple Confidence Rating, (B) the Quadratic Scoring Rule and (C) the mechanism of the Matching Probability .

2.1.1 Simple Confidence Rating

The Simple Confidence Rating (SCR) rule just requires the subject to report her confidence on a numerical scale, without relating any monetary consequence. Nothing is done to provide incentives. The main advantage of such a rule is of course its simplicity.

We implement the SCR as follows. Subjects just have to choose a level of confidence between 0 and 100 (with steps of 5) on a gauge (see Figure 1A). They are told they are free to use the gauge as they want, either by trying to express their confidence level in terms of chance percentages or simply by being consistent in their report with small values for low confidence and high values for high confidence. Payments are independent of elicited probabilities.

2.1.2. The Quadratic Scoring Rule

The Quadratic Scoring Rule (QSR) has been introduced in the 1950's (Brier, 1950) and is extensively used in experimental economics (Nyarko & Schotter, 2002, Palfrey & Wang, 2009) and meteorology (Palmer & Hagedorn, 2006) among others. It is a generalization of the Post-Decision Wagering method (Persaud et al., 2007), which consists in betting on the accuracy of ones' answer. We consider here a very simple version of the QSR. Assume a subject reports a confidence level equal to p. She will then win $(a-b)(1-p^2)$ € if her answer is accurate, and $(a-b)(1-(1-p)^2)$ € otherwise, where a and b are positive constants. The QSR, like the Post-Wagering Method, provides incentives to reveal ones' true confidence (see Gneiting & Raftery, 2007, for a review of proper scoring rules). It is however contaminated by subjects' attitude towards risk (see Offerman et al., 2009, and Dienes & Seth, 2010).

In our experiment, QSR is implemented as follows. We ask subjects to choose among different levels of remunerations (Figure 1B). Each letter corresponds to a payment scheme (x, y), that yields x if their answer is correct and y otherwise. These payments are generated using a QSR with parameters $a=b=10$, and a 0.05 step (i.e., *A* corresponds to $p=1$, *B* corresponds to $p=0.95$ and so on). If, for instance, the subject enters *K*, she will obtain a sure payment of 7.5, which is the optimal choice if she maximizes her expected income and believes that she has an equal probability of being correct or not. The unit used for payments are euro cents. Note that there is no explicit reference to probabilities in this procedure. Subjects are not told that payment schemes are linked to confidence levels.

2.1.3 Matching Probabilities

The third elicitation rule we consider is the Matching Probabilities (MP). It is a variant of the Becker-DeGroot-Marshak mechanism (Becker et al., 1964), and generalizes the no loss gambling method proposed by Dienes & Seth (2010). To elicit a subject's subjective probability about an event E, the subject is asked to provide the probability p that makes her indifferent between a lottery $L(E)$ that gives a positive reward x if E happens, and 0 otherwise and a lottery $L(p)$ that yields a positive reward x with probability p, and 0 with probability $(1-p)$. A random number q is then drawn in the interval [0,1]. If q is smaller than p, the subject is paid according to the lottery $L(E)$. Otherwise, the subject is paid according to a lottery $L(q)$ that yields x with probability q and 0 with probability $(1-q)$.

The no-loss gambling method proposed by Dienes and Seth (2010) is a particular case of the MP. Dienes and Seth were interested in deciding whether subject's confidence is equal to, or higher than, 50%. The method they propose essentially works as follows. If the subject provides a probability higher than 50%, she is paid according to her answer in the type 1 task. If she reports a probability equal or smaller than 50%, she is paid according to a 50-50 lottery. This corresponds exactly to the MP under the two following conditions: (i) the subject can only report two confidence levels: "low" (i.e., 50% or below) or "high" (i.e., higher than 50%), and (ii) the lottery used to fix q is degenerated, so that $q=0.5$ for sure. In the general case, q needs to be random to prevent subjects to overstate their confidence. This is not needed in the no-loss gambling method because it only allows binary answers.

The MP procedure provides incentives for subjects to reveal their subjective probability truthfully. To make this clear, suppose that the subject thinks her probability of success is p but reports a probability $r \neq p$. First consider the case where $r < p$. The lotteries according to which the subject (given her subjective probabilities) is paid are represented in the following table.

	q < r < p	r < q < p	r < p < q
reports *p*	L(E)	L(E)	L(q)
reports r < p	L(E)	L(q)	L(q)

Similarly, assume that the subject reports $r > p$. Her payments (according to her subjective probabilities) are described in the following table.

	q < p < r	p < q < r	p < r < q
reports *p*	L(E)	L(q)	L(q)
reports r > p	L(E)	L(E)	L(q)

It can be observed that, in any case, the subject obtains a lottery that gives her a higher or equal chance to win x if she reports p instead of r.

A major advantage of the MP is that it provides the subjects incentives to reveal her subjective probabilities truthfully, while not being contaminated by her attitude towards risk (see Dienes & Seth, 2010). The MP mechanism might seem complicated. It is thus of interest to investigate whether individuals are able to efficiently use it. As we will see, such is actually the case.

In practice the MP is implemented using a scale of 0 to 100, with steps of 5 (see Figure 1C). After having completed the perceptual task, subjects are told that they are entitled to a ticket for a lottery based on their answers' accuracy. This lottery gives them 0.10€ if their answer is correct, and 0 otherwise. Subjects have then to report on a gauge ranging from 0 to 100 the minimal percentage of chance p they require to accept in an exchange between their lottery ticket and a lottery ticket that gives p chances of winning 0.10€. A number l_1 is drawn according to a uniform distribution between 40 and 100. If l_1 is smaller than p, subjects keep their initial lottery ticket. If l_1 is higher than p, they are paid according to a lottery that gives them l_1 chances of winning. In this case, a random draw determines the payment: a number l_2 is determined using a uniform distribution between 0 and 100, the lottery leads to a win if l_1 is higher than l_2.

2.3. Procedure

2.3.1. Participants

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. They 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 was conducted on computers with 1024x768 screens. We ran two sessions for each rule. We collected in this way data for 35 to 40 subjects for each rule.[1]

2.3.2. Stimuli

The perceptual task we use is a two-alternative forced choice which is known to be a convenient paradigm for SDT analysis (see, e.g., Bogacz et al., 2006). Subjects have to compare the number of dots contained in two circles (see Figure 2A).

The two circles are only displayed for a short fraction of time, about one second, so that it is not possible to count the dots. Subjects have to tell which circle contains the higher number of dots. We allow the difficulty of the task to vary, by changing the spread of the number of dots between the two circles. One of the two circles always contains 50 dots. Its position (to the

[1] These experiments were part of Sébastien Massoni's Master's thesis (see Massoni, 2009).

left or the right of the screen) is randomly chosen for each trial. The other circle contains 50 ± α_j dots, where α_j is randomly chosen for each trial in the set $\{\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4\}$; for all subjects, $\alpha_0=0$ and $\alpha_4=25$. The intermediate difficulty levels are adapted to each participant, in order to control for differences in individual abilities. During a training part of the experiment, α_2 is adjusted so that the subject succeeds in 71% of the cases at that level of difficulty. This calibration is done by using a one-up two-down psychophysics staircase (Levitt, 1971). The two other parameters α_1 and α_2 are then given by $\alpha_3=2.\alpha_2$ and $\alpha_1=\alpha_2/2$ if α_2 is even, and $\alpha_1= (\alpha_2+1)/2$ if α_2 is odd.

This task was completed by a quiz with questions related to logic and general knowledge that is not used in the present study.

2.3.3. Procedure

In a given experimental session, a single elicitation rule (the same for all subjects) is used. Thus, our study is based on a between-subjects analysis with a simple 3x1 design.[2] After the instructions (that include a detailed presentation of the elicitation rule) and a short questionnaire, the experiment is divided in three parts.

In the first part of the experiment, subjects have to answer a randomly chosen quiz (logic or general knowledge) and to provide their confidence for each answer. They have no feedback on their answers.

During the second part of the experiment, subjects have to perform the perceptual task. They begin with a training phase during which the difficulty of the task is calibrated. Confidence is not elicited during this first phase, and they get feedback on their success after each trial. Subjects then perform 100 trials of the perceptual task, and provide their confidence in their answer for each trial. They get feedback on their success in the task and the accuracy of their reported confidence. Furthermore, every each 10 trials, subjects receive a summary of their performance in the last ten trials in terms of success rate and cumulated gains.

The last part of the experiment is similar to the first one, except that subjects have to answer the quiz that has not been selected in the first part.

[2] Pilot experiments have shown that subjects get confused when asked to use different elicitation rules in the same experimental session.

The payment contains three parts. There is a show-up fee of 5€. Subjects are paid for each trial. For groups using the QSR or the Matching probabilities, each 100 trials of the perceptive task is rewarded according to the elicitation rule used, with a maximum payment of 0.10€ and a minimum of 0€. Subjects in the group using the Simple Confidence Rating are paid 0.10€ for each correct answer. Subjects are also paid for the quiz task, but this payment is totally independent.

2.4. Analysis

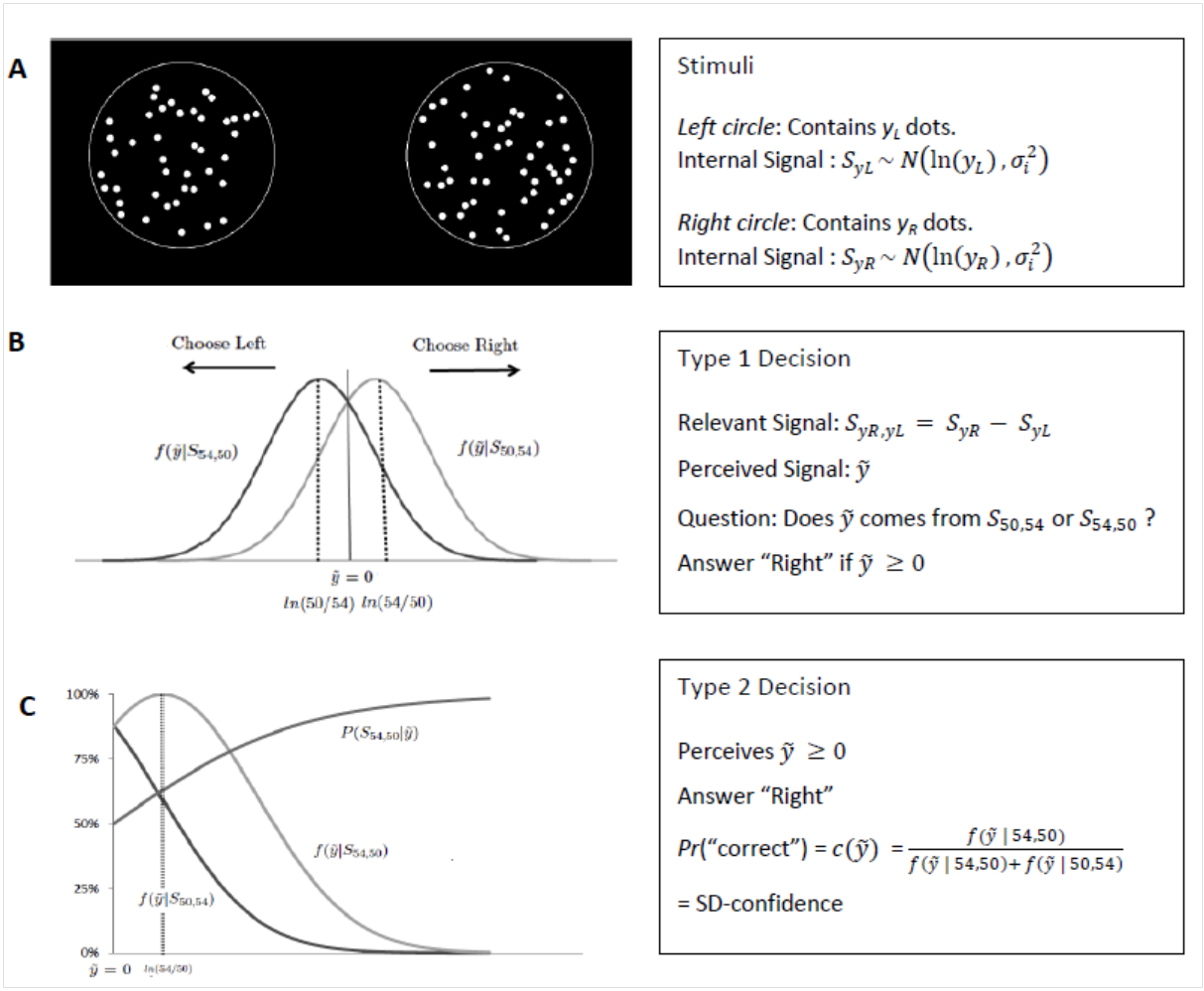


Figure 2: SDT framework. (A) presents an example of stimuli used for the task and details how the visual signal are coded by SDT. (B) defines the optimal criterion of the type 1 decision, while (C) extends SDT to type 2 decision with the computation of SD-confidence.

2.4.1 SDT for perceptive tasks

Since Green & Swets (1966)'s classical book, SDT has been routinely and successfully used in experimental psychology to study individual decisions in perceptual tasks. Let us apply it to simple perception we used in our experiment (see Section 2.3.2). The two circles are only displayed for a short fraction of time, about one second, so that it is not possible to count the dots. However, the subject is aware that a circle can only contain 54 or 50 dots, and that there is an equal probability for each circle to be the one with the largest number of dots.

It is postulated that stimuli are perceived as noisy signals by the sensory system. Here, we are interested in the numerosity of the circles, i.e., the number of dots they contain. It is assumed that, when presented with a circle that contains y dots, the sensory system actually observes the realization of a random signal y that is distributed according to a Gaussian law, with mean $\ln(y)$ and variance σ_i^2, where σ_i is a parameter describing the degree of precision of the internal representation of numerosity in the brain. When observing two circles with respectively y_L and y_R dots (where L and R stand for left and right, respectively), the subject thus receives two noisy signals S_{yR} and S_{yL} (see Figure 2A). Because the subject has to decide which circle contains the largest number of dots, the relevant information is actually the *difference* between the two signals. We thus assume that, when presented with the circles and asked which one contains the largest number of dots, the subject's decision is based on a noisy signal $S_{yR,yL} = S_{yR} - S_{yL}$. For a given trial, the subject thus perceives a signal $\tilde{y}$ and has to decide whether it comes from $S_{yR,yL} = S_{54,50}$ (i.e., there are 50 dots in the left circle, and 54 in the right one), or from $S_{yR,yL} = S_{50,54}$ (i.e., there are 54 dots in the left circle, and 50 in the right one). Let $f(\tilde{y} \mid S_{yR,yL})$ be the density function of $\tilde{y}$ conditional to S_{yR} and S_{yL}. Since she is aware that there is an equal chance for any circle to be the one containing the largest number of dots, her optimal strategy (in the sense of Neyman-Pearson) is based on the likelihood ratio and consists in answering "Right" whenever $\tilde{y} \geq 0$, and "Left" otherwise (see Figure 2B and Green & Swets, 1966).

It has been shown that such a model accounts well for individual decisions, in the sense that the proportion of correct answers as a function of the difficulty of the task (i.e., the ratio y_R / y_L) predicted by the model is very close to that actually observed (Pica et al., 2004).

2.4.2. SDT for confidence

The Bayesian reasoning can be pushed further (see Galvin et al., 2003, Fleming & Dolan, 2010, Rounis et al., 2010, Maniscalco & Lau, 2012) to modelize how subjects make confidence judgments in terms of probabilities about their decisions in a perceptive task. Such judgments are known as "type 2 tasks" (Clarke et al., 1959, Pollack, 1959), as opposed to "type 1 tasks" which consist of discriminating between perceptual stimuli. Consider a trial where the subject perceives a positive signal $\tilde{y}$, and therefore answers "Right". Based on the SDT model presented above, one can easily deduce the probability that she gives the correct answer. According to the Bayes rule, it is equal to $P(S_{54,50} \mid \tilde{y}) = \frac{f(\tilde{y} \mid S_{54,50})}{f(\tilde{y} \mid S_{54,50}) + f(\tilde{y} \mid S_{50,54})}$ (see Figure 2C). This confidence based on signal detection will be called SD-confidence (where "SD" stands for "Signal Detection") in the sequel.

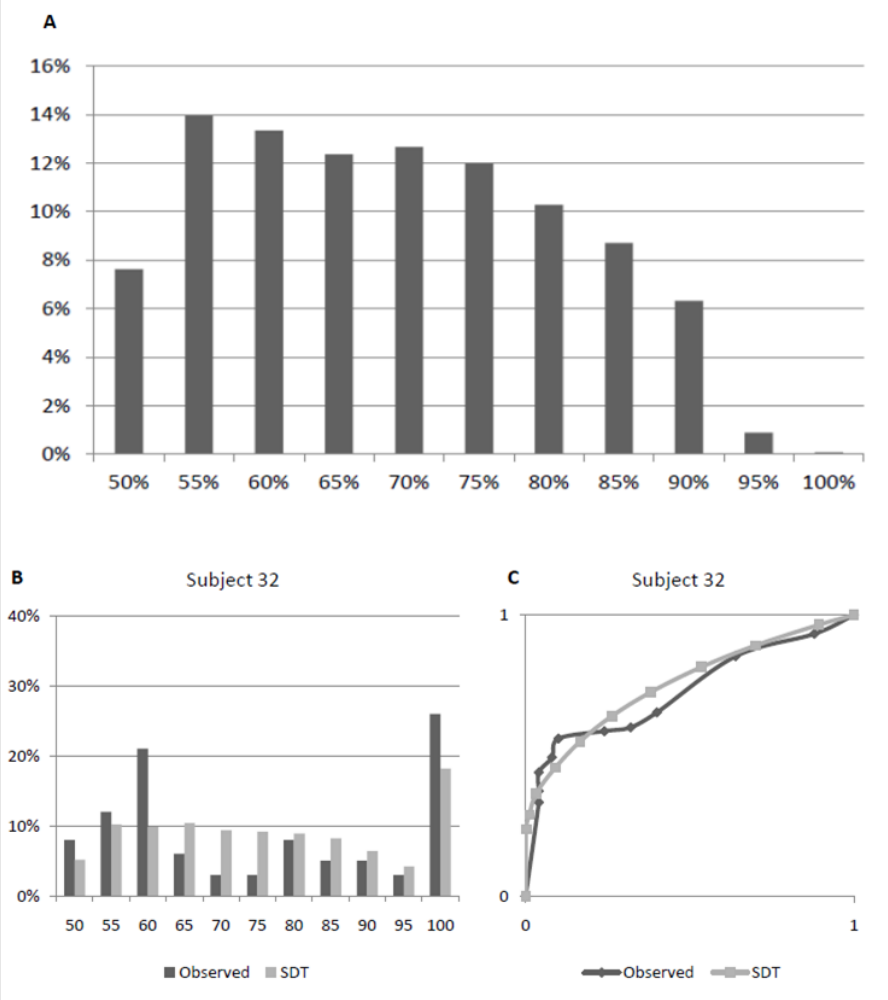


Figure 3: Predictions of confidence. (A) is the predicted distribution of SD-confidence. (B) and (C) are respectively the observed and predicted confidence distribution and AU2ROC for a specific subject.

Since we control for the difficulty levels of the stimuli used in the perceptive task, we can use SDT to estimate subjects' perceptive sensibility from behavioral data (success rates). This leads to an estimation of the distribution of the internal signal used by the subject when performing the perceptual task. With this in hand, the SDT model provides precise predictions about the SD-confidence levels of an ideal (i.e., optimal and Bayesian) observer who receives the same internal signals as the subject. First we can compute the *distribution* of SD-confidence. Indeed, SDT predicts the SD-confidence level associated to each level of the internal signal (Figure 2C). It also provides the probability to reach any confidence level. Given a probability p, $\tilde{y}_p$ is such that $P\left(S_{54,50} \middle| \tilde{y}_p\right) = p$. The probability of observing a confidence level above p is $\int_{\tilde{y}_p}^{\infty} \left(0.5.f\left(\tilde{y} \mid S_{54,50}\right) + 0.5.f\left(\tilde{y} \mid S_{50,54}\right)\right) d\tilde{y}$. In our experiment where the confidence scale is discrete with 5% increments, we can thus deduce the probability distribution of SD-confidence (Figure 3A).

One drawback of the distribution of SD-confidence is that it does not keep track of the trial-by-trial relationship between SD-confidence and success in the perceptive task. This link can be represented by a Receiver Operating Characteristic (ROC) curve (Green & Swets, 1966). Consider a given level of SD-confidence, say 70%. Assume that one uses this confidence level to decide whether the answer was correct or not. Thus, all trials for which the SD-confidence is higher than 70% will be classified as correct, whereas the others will be classified as incorrect. This classification is of course imperfect. But we can compute the false alarm rate (i.e., the proportion of trials that would be wrongly classified as correct) and the hit rate (i.e., the proportion of trials that would be correctly classified as correct). Thus, for each SD-confidence level, we can associate a point on a graph with hit rates on the vertical axis, and false alarm rates on the horizontal axis. The curve that relates all the points obtained by varying the SD-confidence level is the type 2 ROC curve. To measure how accurate confidence is predictive of success, one usually computes the area under this ROC curve (AU2ROC) which has the following statistical meaning. Consider a situation in which trials are already correctly classified into two groups (success and failure) and pick randomly a pair

of trials, one from each group. The probability that the trial with the higher confidence comes from the success group is equal to the AU2ROC.

To illustrate this method, we computed the distribution of elicited confidence and predicted SD-confidence (Figure 3B) for a subject in our experiment. It can be observed that data fit SDT predictions nicely. We also computed, for the same subject, the observed and predicted type 2 ROC curve (Figure 3C). The predicted AU2ROC is equal to 0.75, which is very close to the observed AU2ROC (equal to 0.72). Note that the shape of confidence distribution for this subject differs from that shown in Figure 3A. This is due to the fact that the level of difficulty of the task is not constant in our experiment.

In terms of behaviour, the main prediction of the SDT model described above is a positive relationship between type 1 and type 2 performances. Studies in humans (Maniscalco & Lau, 2012), rhesus monkeys (Kiani & Shalden, 2009) and rats (Kepec et al., 2008) indeed found such a relationship, although it has also been shown that, in some circumstances (e.g., subliminal stimuli) type 1 and type 2 performances might be disconnected (see, e.g., Kanai et al., 2010).

If it is assumed that the elicitation rule leads individuals to report their SD-confidence, then subjects should report confidence levels close to those predicted by SDT. Therefore, the distribution of elicited confidence and the elicited type 2 ROC should be close to that predicted by SDT. Moreover, elicited type 2 ROC could never be better than the one predicted, i.e., the elicited AU2ROC should not be greater than the one predicted. Indeed, predicted confidence levels are those of a perfect bayesian observer, and the subject could therefore not do better (provided the SD-confidence model holds, naturally). Furthermore, if a subject is a good (respectively, bad) assessor of her own SD-confidence, then both the distribution of elicited confidence and the type 2 ROC should be close to (respectively, distant from) the predicted ones. Thus, distances to predicted distribution of confidence and predicted AU2ROC should be positively correlated. Finally, because SD-confidence and the perceptive task are based on the same signals one should observe a positive correlation between performance in the perceptive task and elicited AU2ROC. We summarize these predictions for future reference. A good elicitation rule of SD-confidence should provide:

- *Prediction 1*: an elicited confidence close to predicted SD-confidence;

- *Prediction 2*: an elicited AU2ROC close to the predicted one;

- *Prediction 3*: an elicited AU2ROC not greater than the predicted one;

- *Prediction 4*: the closer the elicited confidence distribution is to the predicted SD-confidence distribution, the closer the elicited AU2ROC is to the predicted one;

- *Prediction 5*: a positive correlation between performance in perceptive task and elicited AU2ROC.

2.4.3. Statistical Tools

The relationships between different measures were analyzed with Pearson's product-moment correlations. Comparisons of their means were conducted using paired t-tests. Measures of distance between two distributions are based on Chi-Square metric and between two cumulative distributions are based on Kolmogorov-Smirnov metric.

Finally, in order to measure the distances between observed and predicted distributions of confidence we construct a measure, called below *ROC_distance*, as follow:

$$ROC\ distance = \frac{|Predicted\ AU2ROC - Observed\ AU2ROC|}{Predicted\ AU2ROC - 0.5}.$$

3. Results

3.1. Elicited Confidence: Descriptive analysis

We start by presenting some basic facts concerning elicited confidence. First, we observe that while the cumulative distributions of elicited confidence obtained by the SCR and the MP are similar, the one corresponding to the QSR differs significantly (see Figure 4A). The difference is mainly due to the fact that the confidence levels elicited by the QSR are strongly concentrated on two values, 50% and 100%. Almost two thirds of elicited probabilities are either equal to 50% or 100% when the QSR is used, which is twice as much as for the two other rules.

Let us next have a look at how subjects' stated confidence is related to their actual success rate (see Figure 4B). A first observation is that, whatever the elicitation rule used, subjects are

generally overconfident. Moreover, the difference between stated confidence and observed success rates increases with stated confidence. If we consider all trials (for both tasks) for which subjects reported a 100% probability of success, we observe an actual success rate of about 84% only. On the other hand, low confidence levels (around 50%) correspond to actual success rates that are slightly higher than 50%. Finally, we note that none of the elicitation rules provides a strictly increasing relationship between stated confidence and the actual success rate.

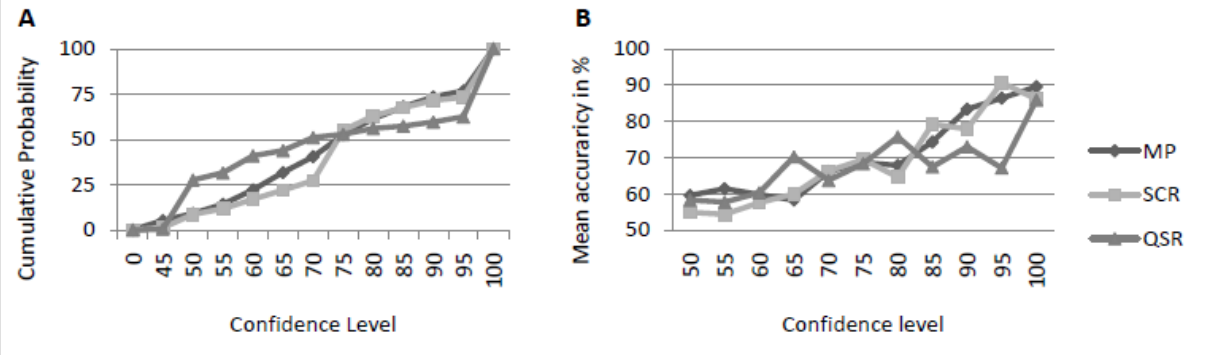


Figure 4: Stated values of confidence. (A) shows the cumulative probability distribution of stated confidence levels for the three rules. (B) represents the link between stated confidence and mean level of accuracy for the three rules.

The QSR and the MP are cognitively demanding and we expect their performances to increase with practice. Our experiment is designed so as to offer subjects the opportunity of learning by using feedback. During the second part of the experiment, subjects used 100 times the elicitation rule with feedback. They could thus have learnt to use the elicitation rule during this part. We can therefore measure learning effects by comparing results for the first half (first 50 trials) and the second half (50 last trials) of the perceptive task. Overall we observe a learning effect for discrimination ability: the AU2ROC is systematically higher in the second part (mean 0.6573, s.d. 0.09) than in the first part (mean 0.6729, s.d. 0.09, $t(113) = -1.8472$, $P = 0.0337$). Nevertheless this learning effect is too weak to be observed at the rule level (for MP: $t(40) = 0.8814$, $P = 0.1918$; for QSR: $t(35) = 1.3079$, $P = 0.0998$; and for SCR: $t(38) = 0.9935$, $P = 0.1635$). Since the increase is quite similar for the three rules, it is likely that this learning effect reflects more an increase in metacognitive abilities than an increase in the understanding of the QSR and the MP.

3.2. SD-Confidence

We now consider to what extend elicitation rules lead individuals to report confidence levels that are close to those predicted by the SD-Confidence model. The first thing we need is to compute predicted confidence in the perceptive task. The only problem here is that there are actually five levels of difficulty. We extend the Bayesian analysis described in Section 2.4.2 to this case, under the assumption that subjects have correct priors on the distribution of difficulty levels.

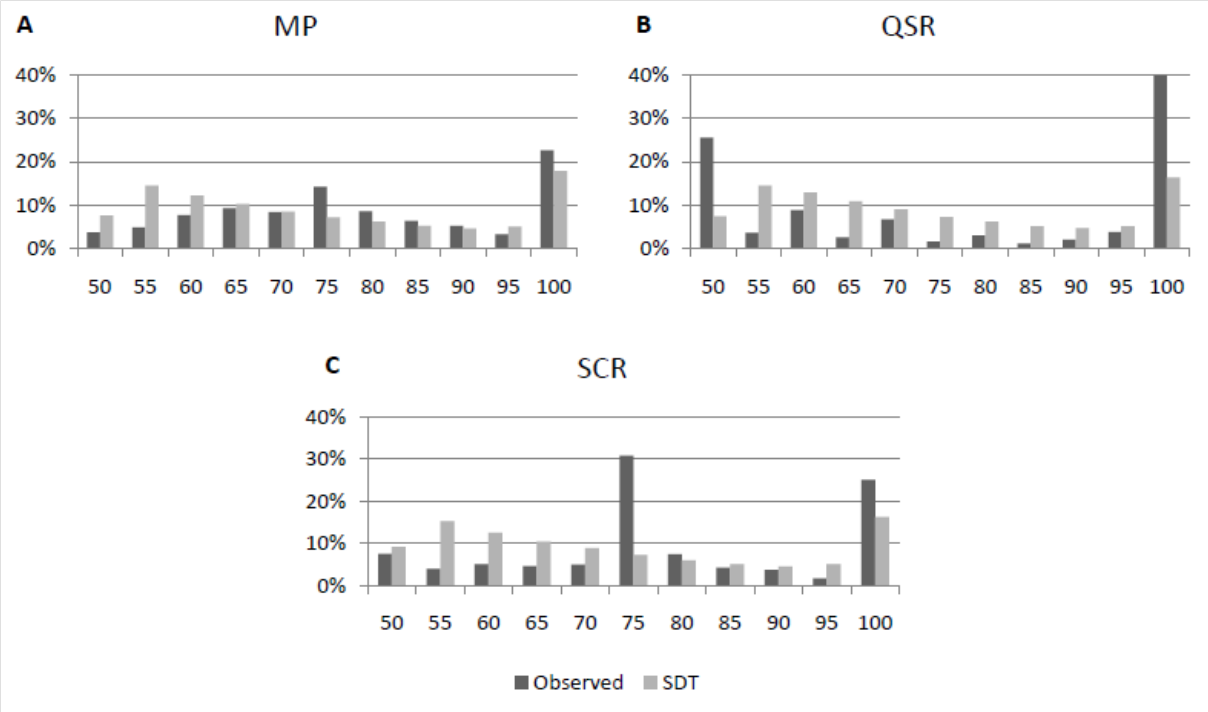


Figure 5: Observed and predicted distribution of confidence. (A), (B) and (C) are respectively the observed and predicted distribution of confidence for MP, QSR and SCR.

We proceed by examining in turn each of the predictions 1 to 5 listed at the end of Section 2.4.2. Let us start with prediction 1, which states that elicited confidence should be close to predicted SD-confidence. A first answer is given by comparing elicited confidence and predicted SD-confidence distributions. Figure 5 reports the elicited confidence and predicted SD-confidence distributions for each elicitation rule (data are pooled across all levels of difficulty and all subjects). It appears clearly that the MP is the rule that leads to the best fit. The SCR is plagued by the large proportion of elicited confidence levels equal to 75%, which

is the pre-filled value of the gauge[3]. Confidence levels elicited with the QSR are those that differ the most from predicted SD-confidence. There is a peak at a 50% confidence level, which is expected because of risk aversion. But we also observe a high peak at the 100% value (with 39.9% of the answers), which cannot be explained by risk aversion, and which does not correspond to predictions of SDT (only 18% of the answer should take this value according to SDT).

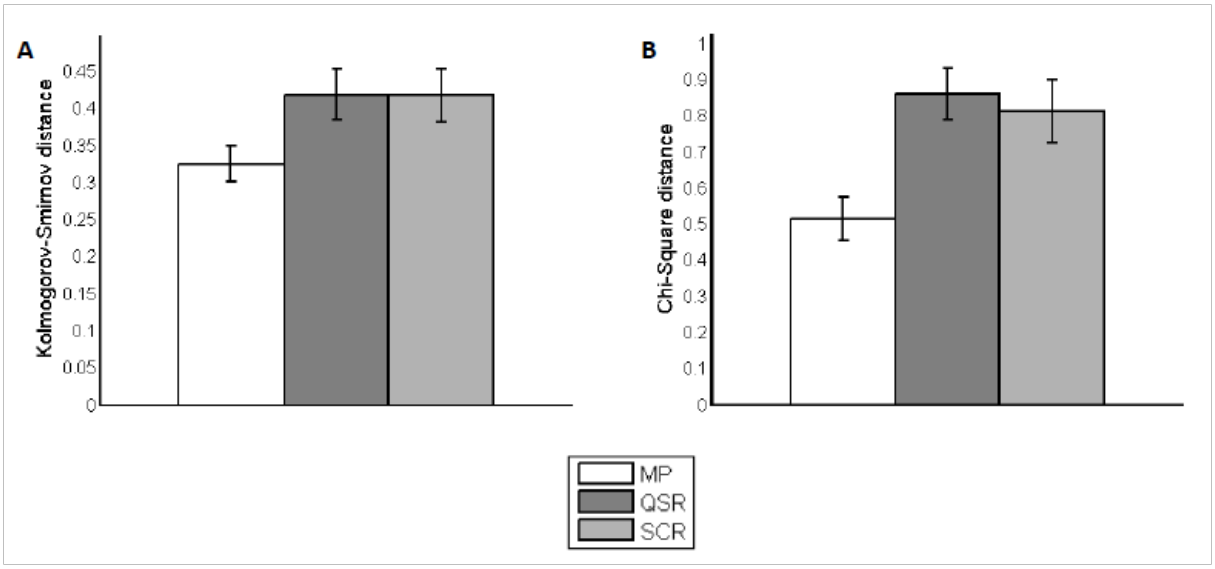


Figure 6: Distance between confidence distributions. (A) presents the Kolmogorov-Smirnov distance between the cumulative distribution of stated and predicted confidence for the three rules. (B) shows the Chi-Square distance between observed and predicted confidence distribution of the three rules.

To confirm the visual impression that MP leads to the best fit between elicited confidence and predicted SD-confidence, we computed the Chi-Square distance between the elicited confidence and predicted SD-confidence distributions, and the Kolmogorov-Smirnov (KS) distance between the elicited confidence and predicted SD-confidence cumulative distributions. We report the two distances for the three rules (with s.d. in brackets) in Figure 6. The results for t-tests show that the two distances are significantly lower for the MP (Chi-Square distance: mean 0.5152, s.d. 0.37; KS distance: mean 0.3252, s.d. 0.15) than for the QSR (Chi-Square distance: mean 0.8621, s.d. 0.42, t(73) = -3.8131, P = 0.0001; KS distance: mean 0.4182, s.d. 0.20, t(73) = -2.3281, P = 0.0113) and the SCR (Chi-Square distance: mean 0.8129, s.d. 0.52, t(75) = -2.9005, P = 0.0024; the KS distance: mean 0.4158, s.d. 0.22, t(75) = -2.2143, P = 0.0149) while there are no significant differences between QSR and SCR (Chi-

[3] We do not observe such a result for the MP that is also based on a gauge pre-filled at 75%. We suspect that this is due to the fact that no incentive is provided in the SCR, and that this might lead subjects simply not to make the effort to change this value in many cases.

Square distance: t(70) = 0.4405, P = 0.3305; KS distance: t(71) = 0.0096, P = 0.4962). We also found that the two distances are strongly correlated (r = 0.85, P < 0.00001).

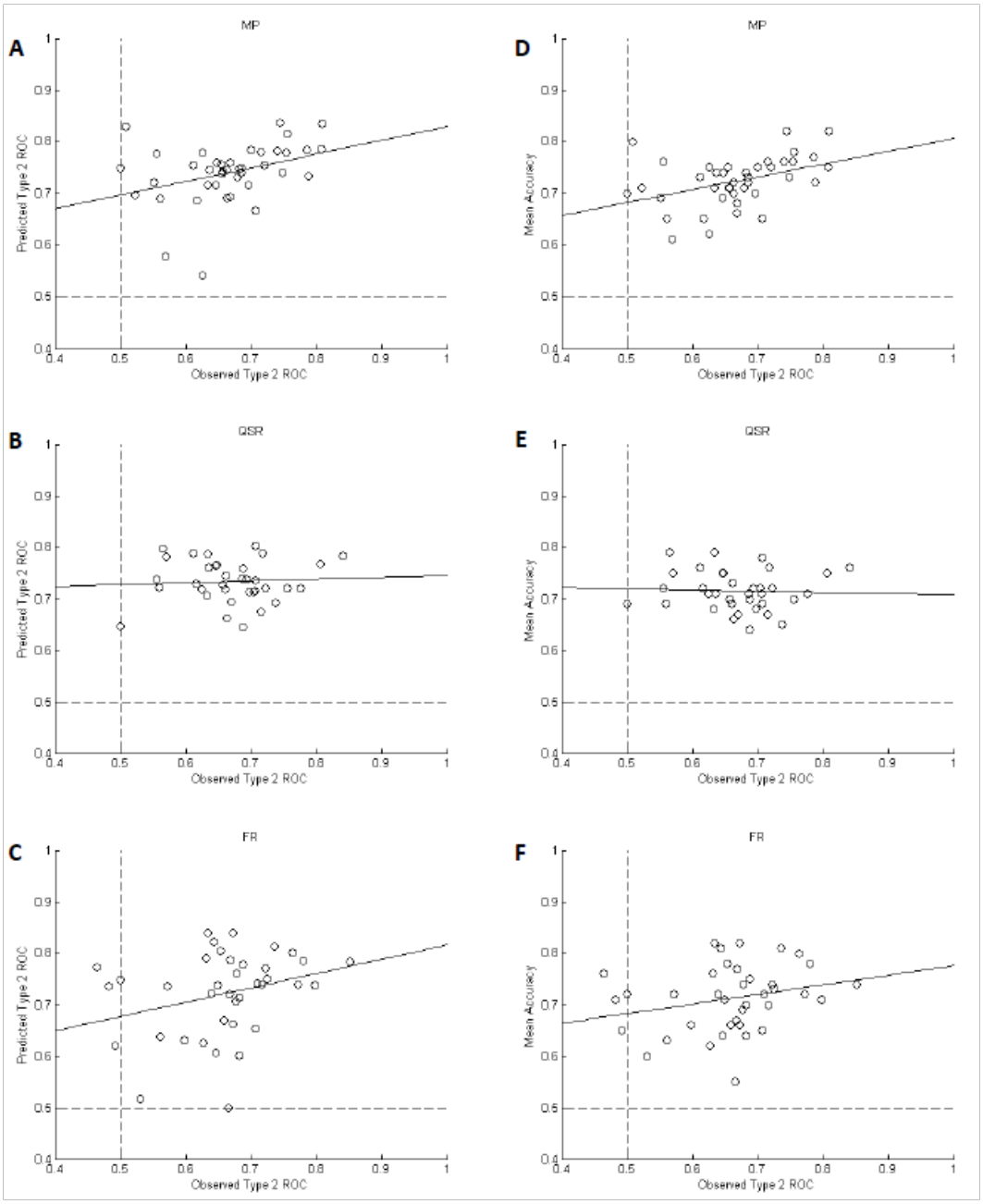


Figure 7: Correlations between AU2ROC and accuracy. (A), (B) and (C) show respectively the correlations between observed and predicted AU2ROC for the three rules; while (D), (E) and (F) present respectively the correlations between observed AU2ROC and mean levels of accuracy for the three rules.

The second prediction states that elicited AU2ROC should be close to predicted ones. Figure 7 (panels A, B and C) displays the corresponding data for each rule. The correlation between observed and predicted AU2ROC is positive and statistically significant for the MP (Fig. 7A,

r = 0. 36, P = 0.0232) and for the SCR (Fig. 7C, r = 0.29, P = 0.0741) while it is not statistically significant for the QSR (Fig. 7B, r = 0.06, P = 0.7233).

Our third prediction is that observed AU2ROC should not be greater than the predicted one. This is actually the case for 34 out of 40 subjects (85%) in the MP group, 28 out of 38 (74%) in the SCR group and 26 out of 35 (74%) in the QSR group.

If elicited confidence corresponds to SD-confidence, then a good (respectively, bad) elicitation rule should be good (respectively, bad), for both the distribution of confidence and the type 2 ROC (in the sense of giving results close to those predicted by SDT). This is our fourth prediction. In other words, we should observe a positive correlation between the distance between observed and predicted confidence distributions on the one hand, and the distance between observed and predicted AU2ROC on the other hand. As an indicator of distance between observed and predicted AU2ROC we use the variable *ROC_distance*. The correlations are positive and significant for the MP (with the Chi-square distance: r = 0.48, P = 0.0016; with the KS distance: r = 0.52, P = 0.0007) and the SCR (with the Chi-square distance: r = 0.60, P = 0.0001; with the KS distance: r = 0.36, P = 0.0289). In contrast, the results are less conclusive for the QSR, for which we observe a correlation between distances measured by the KS metric (r = 0.49, P = 0.0030) but not by the Chi-square metric (r = 0.19, P = 0. 2828).

Our last prediction concerning SD-confidence is that we should observe a positive correlation between the mean success rate in the type 1 task and the observed AU2ROC. We report these correlations in Figure 7 (panels D, E, F). We found that performances in type 1 and type 2 tasks are strongly correlated when confidence is elicited with MP (Fig. 7D, r = 0.41, P = 0.0086). The correlation is still positive, but not significant for the SCR (Fig. 7F, r = 0.25, P = 0.1271). More strikingly, we found no correlation between performances in type 1 and type 2 tasks when the QSR is used (Fig. 7E, r = -0.04, P = 0.8004).

Taken together, our results suggest that elicitation rules differ strongly in the kind of confidence they convey. Whereas confidence levels reported using MP are globally compatible with predicted SD-confidence, those obtained through QSR can hardly be explained by the classical SDT model. The results concerning the SCR are less conclusive. Our conclusion at this point should thus be that MP seems a good rule (compared to the other ones), if one seeks to elicit SD-confidence.

5. Discussion

Dienes and Seth (2010) compared three methods for measuring consciousness: verbal report, post-wagering method, and an original "no-loss gambling" procedure. They found that, in an implicit learning task (artificial grammar), the no-loss gambling method proved to be no less sensitive than the two other ones, while being immune to subjects' attitude towards risk.

The purpose of this study was similar in spirit. Our aim was to compare different method to measure confidence. We compared three elicitation methods for retrospective confidence judgements in a perceptive task with respect to their ability to fit SDT predictions. We found that the Matching Probability (which is a fine-grained version of the no-loss gambling method) outperforms the Simple Confidence Rating (a direct report on a numerical scale) and the Quadratic Scoring Rule (a post-wagering method). These results thus show that the choice of the method for confidence measurement matters greatly, and provide support for the use of the Matching Probability in studies of confidence judgements that are based on SDT analysis.

A possible explanation for these results could be as follows. First, it is known that QSR is plagued with individuals' risk aversion. Furthermore, it is expressed in terms of stakes, and not in terms of probabilities or confidence levels. It might be that this simple fact requires some "translations" (from probabilities to stakes) that distort individuals' reports. By contrast, both SCR and MP are directly expressed in terms of confidence rating, and are immune to risk aversion. Moreover, MP provides incentives to truthfully report one's confidence, which is not the case of SCR. This might explain that MP performs better.

Incidentally, if one is willing to interpret confidence as a degree of consciousness, our results can also be read as a confirmation of those of Dienes and Seth (2010) for perceptive tasks, fine-grained measurement methods, and using a different comparison criterion (proximity to the prediction of an ideal Bayesian observer).

This study presents some limitations. First we use the most basic SDT model in order to predict SD-confidence. Even if our results are robust enough to draw some conclusion about the ability of the PM to fit SDT predictions, it could be interesting to confirm these results by using SDT in a more sophisticated way. Remaining in a static framework, one could refine the SDT model in order to take into account possible position bias and unequal variance of

signals (Wickens, 2002). If we were to extend our analysis to a dynamic setting, the diffusion model seems to be a powerful tool to understand type 1 (see Ratcliff & McKoon, 2008, for a review) and type 2 (Ratcliff & Starns, 2009, Pleskac & Busemeyer, 2010) decisions. Unfortunately our data does not allow trying these two refinements of SDT. The second drawback of this study comes from our experimental design. Our analysis is based on between-subjects comparison: Each individual only uses one of the three elicitation rules. As metacognitive ability is known to be very heterogeneous between subjects (Fleming et al., 2010) and as a switch of rules during an experimental session has proved to be too confusing, a proper protocol could be to ask subjects to come for several sessions, spaced out by time, with the use of a new rule at each session.

Recent studies on metacognition have mainly focussed on the measure of metacognitive ability and its variation across individuals or across tasks, using SDT analysis as a theoretical framework. In the present study we take another point of view by trying to identify which elicitation method is the most appropriate to measure confidence in line with SDT framework. Our results support the idea that the choice of elicitation rules matters and provide evidence that experiments which use SDT as a theoretical basis should elicit confidence by the Matching Probability mechanism.

Acknowledgments

This research was supported by two grants from the ANR (Riskemotion – ANR-08-RISKNAT-007-01 and Feeling of Control – BLAN-07-2-192879).

References

Becker, G.M., DeGroot, M.H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science, 9 (3)*, 226–32.

Brainard, D.H. (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.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J.D. (2006). The physic of optimal decision making: a formal analysis of models of performance in two-alternative forced choice tasks. *Psychological Review, 113(4)*, 700-765.

Clarke, F., Birdsall, T., & Tanner, W. (1959). Two types of ROC curves and definition of parameters. *The Journal of Acoustical Society of America, 31(5)*, 629-630.

Dienes, Z. (2007). Subjective measures of unconscious knowledge. *Progress in Brain Research*, *168*, 49-269.

Dienes, Z., & Seth, A. (2010). Gambling on the unconscious: a comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition, 19(2),* 674-681.

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Thousand Oaks, CA: SAGE Publications.

Fleming, S.M., & Dolan, R.J. (2010). Effect of loss aversion on post-decision wagering: implications for measures of awareness. *Consciousness and Cognition, 19(1),* 352-363.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541-1543.

Galvin, S., Podd, J., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10(4)*, 843-876.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, *51*(7), 771-781.

Gneiting, T., & Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102(477)*, 359-378.

Green, D.A., & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*. New-York, NY: John Wiley and Sons.

Kadane, J.B., & Winkler, R.L. (1988). Separating probability elicitation from utilities. *Journal of the American Statistical Association, 83(402)*, 357-363.

Kanai, R., Walsh, V., & Tseng, C. (2010). Subjective discriminability of invisibility: a framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition, 19(4)*, 1045-1057.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322-1337.

Kepecs, A., Uchida, N., Zariwala, H., & Mainen, Z. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455(7210)*, 227-231.

Kiani, R., & Shalden, M. (2009). Representation of confidence associated with a decision by neurons in parietal cortex. *Science, 324(5928)*, 759-764.

Levitt, H. (1971).Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49(2)*, 467-477.

Maniscalco, B., & Lau, H (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21(1)*, 422-430

Massoni, S. (2009). A direct revelation mechanism for elicitating confidence in perceptual and cognitive tasks. Master's Thesis, Université Paris 1 – Panthéon-Sorbonne.

McCurdy, L.Y., Maniscalco, B., Metcalfe, J., Liu, K.Y., de Lange, F.P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *The Journal of Neuroscience, 33(5)*, 1897-1906.

Middlebrooks, P. G., & Sommer, M. A. (2011). Metacognition in monkeys during an oculomotor task. *Journal of Experimental Psychology-Learning Memory and Cognition*, *37*(2), 325.

Nyarko, Y., & Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica, 70(3)*, 971-1005.

Offerman, T., Sonnemans, J., Van de Kuilen, G., & Wakker, P.P. (2009). A truth-serum for non-bayesian: correction proper scoring rules for risk attitudes. *Review of Economic Studies, 76(4)*, 1461-1489.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1287-1296.

Palfrey, T., & Wang, S. (2009). On eliciting beliefs in strategic games. *Journal of Economic Behavior and Organization, 71(2)*, 98-109.

Palmer, T.N., & Hagedorn, R. (2006, Eds.). *Predictability of Weather and Climate*. Cambridge, UK: Cambridge University Press.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience, 10(2)*, 257-261.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in Amazonian indigene group. *Science, 306(5695)*, 499-503.

Pleskac, T.J., & Busemyer, J.R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117(3)*, 864-901.

Pollack, I. (1959). On indices of signal and response discriminability. The *Journal of Acoustical Society of America, 31(7)*, 1031-1031.

Ratcliff, R., & McKoon, G. (2008), The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20(4)*, 873-922.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116(1),* 59–83.

Rounis, E., Maniscalco, B., Rothwell, J., Passingham, R., & Lau, H (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience, 1(3)*, 165–175.

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other?. *Consciousness and Cognition*, *19*(4), 1069-1078.

Smith, J. D., Shields, W. E., & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, *26*(3), 317-339.

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, *20*(4), 1787-1792.

Wickens, T.D. (2002). *Elementary Signal Detection Theory*. New-York, NY: Oxford University Press.