Elicitation of Subjective Probabilities in the Light of Signal Detection Theory [*]

Thibault Gajdos [†] Sébastien Massoni [‡] Jean-Christophe Vergnaud [§]

November 2012

JOB MARKET PAPER

Abstract

Most theories of decision under uncertainty are based on the idea that individuals weight events by some (transformation of) subjective probabilities. Elicitation of these subjective probabilities has thus became a major concern for evaluating and applying these models. In order to do so, one needs (*i*) to be sure that such subjective probabilities actually exist in individuals' minds and (*ii*) to be able to infer these subjective probabilities from behavioral data. It is known from cognitive science that decisions in perceptive tasks are based on probabilities encoded at the neuronal level. Moreover, Signal Detection Theory (SDT) provides a theoretical model, confirmed by experimental neuronal data, that relates these probabilities to behavioral data. We use three different elicitation rules to measure individuals' confidence in a perceptive task, and compare the results with the theoretical predictions based on SDT. We find that subjective probabilities elicited by a specific rule, the Matching Probabilities, fit very well the theoretical predictions. We also show that these results are consistent with those obtained for the same subjects in a non-perceptive task (knowledge and logic quiz). This paves the way for extending our findings to non-perceptive tasks.

[*] Financial supports from ANR grants is acknowledged (Riskemotion - ANR-08-RISKNAT-007-01; Feeling of control - BLAN07-2_192879).

[†] GREQAM, CNRS, Aix-Marseille University. thibault.gajdos@univ-amu.fr.

[‡] Corresponding author: sebastien.massoni@gmail.com. CES - University of Paris 1, Paris School of Economics. Address: 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13 France.

[§] CES - University of Paris 1, CNRS. jean-christophe.vergnaud@univ-paris1.fr.

1 Introduction

Most theories of decision under uncertainty are based on the idea that individuals weight events by some (transformation of) subjective probabilities. Elicitation of these subjective probabilities has thus became a major concern for evaluating and applying decision models. This requires to identify procedures (known as elicitation rules) which can be used to measure as accurately as possible subjective probabilities.

However, decision theory itself cannot help to assess how close elicited probabilities are to subjective probabilities an agent has in mind, because it only provides "as if" models. This essentially results from a methodological constraint. Indeed, the assumption that one can only observe the results of choices, and not the decision processes themselves, is a cornerstone of decision theory. For instance, the celebrated Subjective Expected Utility model only says that people decide as if they were maximizing an expected utility, and not that this is what they actually do. All representation theorems, that form the core of mathematical decision theory, are to be interpreted in that way. Decision theory cannot help to elicit subjective beliefs, because it does not even assume that these subjective probabilities actually exist in the first place. The only thing one can do in this framework is to identify the subjective probabilities that would be compatible with individual choices if agents were acting in conformity with a given theoretical model. In order to avoid this circularity while respecting the methodological constraints of decision theory, one needs (*i*) to be sure that subjective probabilities actually exist in decision makers' mind, and (*ii*) to be able to infer these subjective probabilities from behavioral data. This is precisely what we aim at in this paper.

It is known from cognitive sciences that decisions in perceptive tasks are based on probabilities encoded at the neuronal level, which solves the first question. Moreover, Signal Detection Theory (SDT) provides a theoretical model, confirmed by experimental neuronal data, that relates these probabilities to behavioral data. It is thus possible, in this context, to compare these predicted subjective probabilities to those obtained by using various elicitation procedures. The idea being that, if we are able to find an elicitation procedure that delivers subjective probabilities close to those predicted by SDT, it would be a good candidate to elicit subjective probabilities in other contexts as well.

We choose to study three elicitation rules. Since the most widely used elicitation rule in economics (and other fields, such as meteorology) is the Quadratic Scoring Rule (QSR), it is natural to consider it. Since the common practice in psychology is to use a simple ordinal scales without incentives, our second elicitation rule is the Free Rule (FR) which simply requires the subject to report her confidence, without relating any monetary consequence to stated probabilities. The third elicitation rule we consider is the Matching Probabilities (MP), which is a variant of the famous Becker-DeGroot-Marshak mechanism (Becker, Degroot, and Marschak (1964)). It consists in eliciting an objective probability equivalent to a subjective probability.

We compare the results obtained by these three rules to the subjective probabilities predicted by SDT in a very simple perceptive task. We find that MP very nicely fits the predictions of SDT. We also observed that although this rule might seem complicated at first sight, there is no evidence that subjects had more difficulties using it than the two other rules. Moreover, using a task based on a knowledge quiz for the same subjects, we found evidence suggesting that these results might possibly extend to other, non perceptive, tasks. MP thus appears as a good candidate if one seeks for an incentive-compatible elicitation rule that is not affected by rewards, and remains reasonably simple to use.

The remainder of the paper is organized as follows. In section 2, we show how actual subjective probabilities can be predicted from behavior, using SDT. We also provide a short account of the neurosciences literature supporting the idea that this model is grounded at the neuronal level. In Section 3, we present the three elicitation rules we will put under scrutiny. We will then describe the design of our experiment in Section 4. Section 5 contains the main results of our experiment. We close the article by a brief summary and some results concerning the elicitation rules from the classical point of view of calibration and discrimination as measures of good probability assessors.

2 Confidence and probabilities

Our analysis is based on Signal Detection Theory (SDT). In the first part of this section, we briefly present how SDT is used in psychophysics for analyzing perceptive task, and how it can be extended to study confidence. We then describe some evidence supporting the idea that SDT can in some circumstances provide an accurate description of actual neuronal processes for perceptive task and confidence assessment in these tasks. We conclude by describing empirical predictions regarding elicited confidence that can be made on the basis of SDT.

2.1 Signal Detection Theory as behavioral model

2.1.1 SDT for perceptive tasks

Since Green and Swets (1966)'s classical book, SDT has been routinely and successfully used in experimental psychology to study individual decisions in perceptual tasks. The idea is the following. Consider a simple perceptual task, where subjects have to compare the number of dots contained in two circles (see Figure 1). 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.

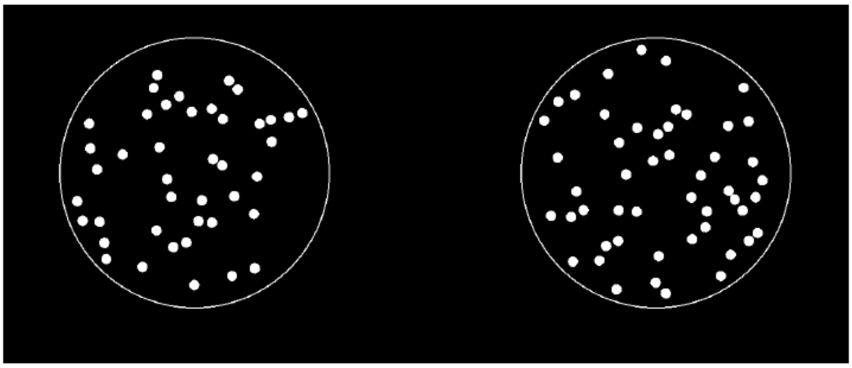

Figure 1: Perceptual task

One postulates 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. One assumes that, when presented with a circle that contains y dots, the sensory system actually observes a realization of a random signal S_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_{y_R} and S_{y_L}. 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 $S_{y_R,y_L} = S_{y_R} - S_{y_L}$.

On a given trial, the subject thus perceives a signal $\tilde{y}$ and has to decide whether it comes from $S_{y_R,y_L} = S_{54,50}$ (i.e., there are 50 dots in the left circle, and 54 in the right one), or from $S_{y_R,y_L} = S_{50,54}$ (i.e., there are 54 dots in the left circle, and 50 in the right one). Denote $f(\tilde{y}|S_{y_R,y_L})$ its density function. 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 based on maximum likelihood consists in answering "Right" whenever $\tilde{y} \geq 0$, and "Left" otherwise (see Figure 2).

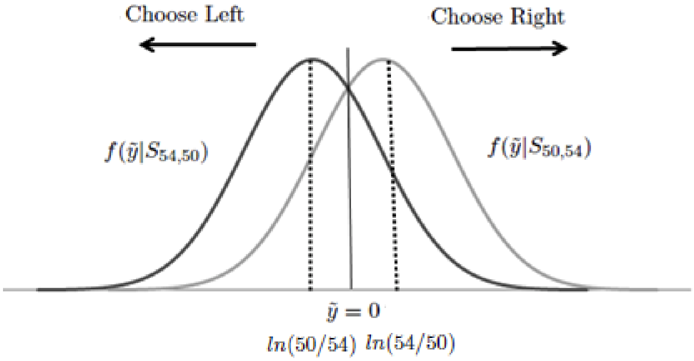

Figure 2: SDT in the two-alternatives forced choice

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, Lemer, Izard, and Dehaene (2004)).

2.1.2 SDT for confidence

The Bayesian reasoning can be pushed further (see Galvin, Podd, Drga, and Whitmore (2003), Fleming and Dolan (2010), Rounis, Maniscalco, Rothwell, Passingham, and Lau (2010), Maniscalco and Lau (2012)) to modelize how subjects make confidence judgments in terms of probabilities on their decisions in a perceptive task. Such judgments are known as "type 2 tasks" (Clarke, Birdsall, and Tanner (1959), Pollack (1959)), as opposed to "type 1 tasks" consisting in discriminating between perceptual stimuli.

Consider a trial where the subject perceives a positive signal $\tilde{y}$, and therefore answer "Right". Based on the SDT model presented above, one can easily deduce the probability that she had given the correct answer. By Bayes rule it is equal to $P(S_{54,50}|\tilde{y}) = f(\tilde{y}|54,50)/(f(\tilde{y}|54,50) + f(\tilde{y}|50,54)$ (see Figure 3). This confidence based on signal detection will be called SD-confidence (where "SD" stands for "Signal Detection") in the sequel.

Since we only collect behavioral data in our experiment, we cannot measure the neuronal firing rate corresponding to the internal signal used by the subject to perform the perceptive task. However, since we control for the difficulty levels of the stimuli used in the perceptual 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

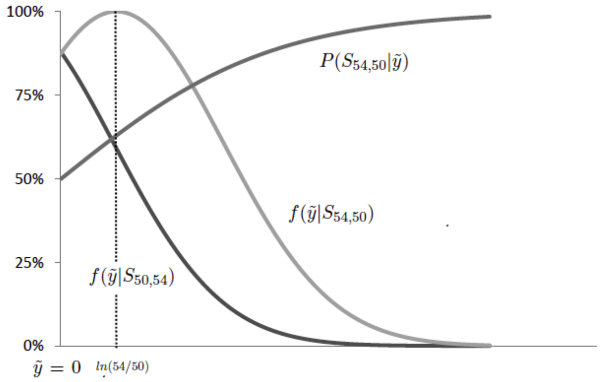

Figure 3: SD-confidence

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 3). It also provides the probability to reach any confidence level. Given a probability p, let $\tilde{y}_p$ be such that $P(S_{54,50}|\tilde{y}_p) = p$. The probability to observe a confidence level above p is $\int_{\tilde{y}_p}(0.5f(\tilde{y}|54,50) + 0.5f(\tilde{y}|50,54))d\tilde{y}$. In our experiment where the confidence scale is discrete with 5% increments, we can thus deduce the probability distribution of SD-confidence (Figure 4).

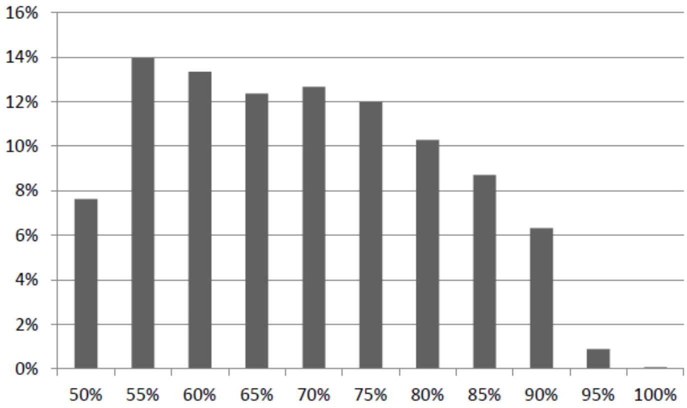

Figure 4: Distribution of SD-confidence

One drawback of the distribution of SD-confidence is that it does not keep tracks of any relationship between SD-confidence and success in the perceptive task. This link can be represented by a Receiver Operating Characteristic (ROC) curve (Green and 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 other will be classified as incorrect. This classification is of course imperfect. But we can precisely 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, to 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 is the one from the success group is equal to the AU2ROC.

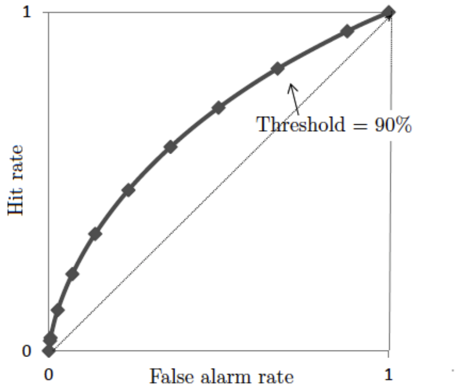

Figure 5: ROC curve

To illustrate, we computed the distribution of elicited confidence and predicted SD-confidence (Figure 6) for a subject of our experiment. One observes that data fit nicely SDT predictions. We also computed, for the same subject, the observed and predicted type 2 ROC curve (Figure 7). 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 4. This is due to the fact that the difficulty level of the task is not constant in our experiment.

In terms of behaviors, the main prediction of the SDT model described above is

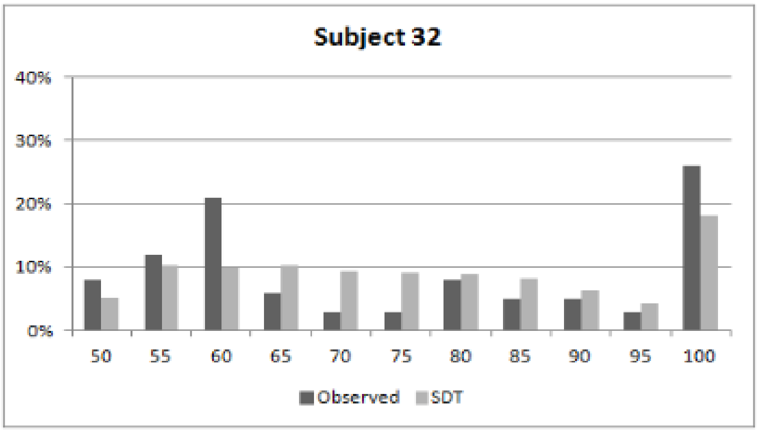

Figure 6: Observed and SDT confidence distribution for one specific subject.

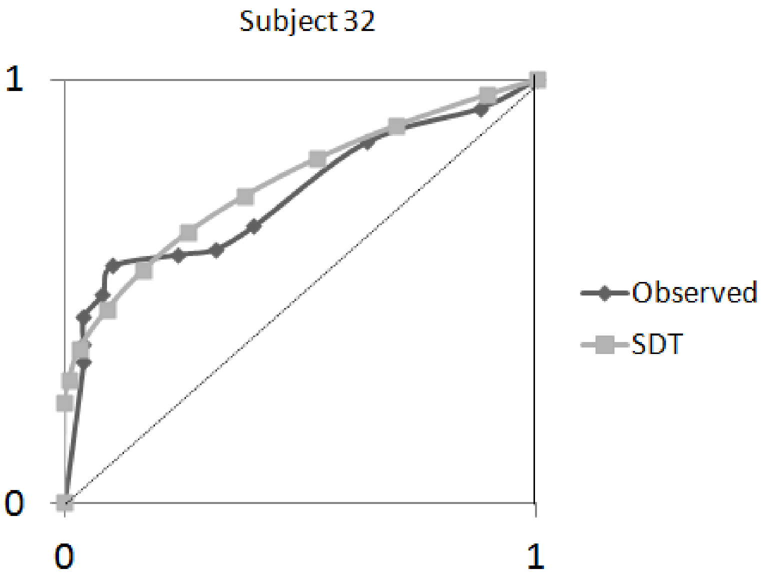

Figure 7: Observed and SDT ROC curve for one specific subject.

a positive relationship between type 1 and type 2 performances. Studies in humans (Maniscalco and Lau (2012)), rhesus monkeys (Kiani and Shadlen (2009)) and rats (Kepecs, Uchida, Zariwala, and Mainen (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, Walsh, and Tseng (2010)).

2.2 Neuronal evidence

The mere fact that SDT accommodates observed behavior in many situations does not imply that it provides an accurate description of the process that actually takes place in the brain. There is, however, a substantial amount of evidence coming from neurosciences that support this idea. We review briefly here some of this evidence.

2.2.1 Perceptual tasks

First, it has been shown that there exist single neurons tuned to numerosity in the macaque monkey, that encode numerical quantities (Nieder, Freedman, and Miller (2002), Nieder and Miller (2004)). Using neuroimaging and psychophysics Piazza, Izard, Pinel, Le Bihan, and Dehaene (2004) established that a similar neural coding scheme is likely to exist for human (see Nieder and Dehaene (2009) for a review). The population of numerosity-selective neurons taken together encodes numerosity into a log-Gaussian distribution, exactly as assumed by SDT (Nieder and Miller (2003)).

Furthermore, neuroscientists have provided empirical and theoretical evidence that populations of neurons in the lateral intraparietal cortex (LIP) can encode the uncertainty about stimuli under the form of probability distributions, and combine this information in a bayesian way (Gold and Shadlen (2002), Ma, Beck, Latham, and Pouget (2006), Yang and Shadlen (2007), Beck, Ma, Kiani, Hanks, Churchland, Roitman, Shadlen, Latham, and Pouget (2008)). In particular, populations of neurons can encode the all posterior probability distribution associated with a stimulus.

Thus, SDT should not simply be considered as an "as if" model of decision making in simple perceptive contexts, but as an elementary description of *how* decisions are actually made. In particular, there is strong and converging evidence that neurons in the LIP area encode a probabilistic information and use it to carry out bayesian inference when performing simple perceptive tasks. This implies in particular that probability distributions involved in such decision processes should be considered as real mental representation used by the subject to make her choices, and not simply as modeling tool of the observer.

2.2.2 SD-Confidence

Studies using neuronal recording in the rhesus monkey (Kiani and Shadlen (2009)) and the rat (Kepecs, Uchida, Zariwala, and Mainen (2008)) show that animals' confidence in their decisions, measured by post-decision choices, can be explained and predicted by SDT. Moreover, Kiani and Shadlen (2009) show that the same neurons are involved in type 1 and type 2 decisions, and thus support the hypothesis that confidence is based on the same signals on which type 1 decisions are made, which is the fundamental prediction of SDT.

However, it has also be shown that there is an heterogeneity across reported confidence (measured by difference between AU2ROC), even when controlling for the individual performance level in the perceptive task. This suggests that confidence reports involve other cognitive processes than type 1 decisions. Confirming previous studies (Rounis, Maniscalco, Rothwell, Passingham, and Lau (2010), Fleming, Weil, Nagy, Dolan, and Rees (2010)), Fleming, Huijgen, and Dolan (2012) show, using fMRI analysis, that a specific brain area (the right rostrolateral prefontal cortex, rlPFC) is involved in confidence judgments. Moreover, they found an increase of connectivity with visual cortex when subjects report their confidence in a visual perceptive task. They conclude that it is likely that signals used for type 1 decisions might be re-represented in specific brain area for confidence reports. Individual heterogeneity in the quality of type 2 ROC may thus be explained by neuro-anatomical differences (Fleming, Weil, Nagy, Dolan, and Rees (2010)).

2.3 Empirical predictions from SDT

Assume that the elicitation rule yields individuals to report their SD-confidence. Then, subjects should report confidence levels close to the one predicted by SDT. Therefore, the distribution of elicited confidence and the elicited type 2 ROC should be close to that predicted by SDT. Moreover, the elicited type 2 ROC could never be better than the predicted one, i.e., elicited AU2ROC should not be greater than predicted one. Furthermore, if a subject is a good (respectively, bad) assessor of her 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 is based on the same signals than those used for the perceptive task, 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 yield:

A1 elicited confidence close to predicted SD-confidence;

A2 elicited AU2ROC close to predicted one;

A3 elicited AU2ROC not greater than predicted one;

A4 the closer is the elicited confidence distribution to the predicted SD-confidence distribution, the closer is the elicited AU2ROC to predicted one;

A5 positive correlation between performance in perceptive task and elicited AU2ROC.

3 Elicitation Rules

The main objective of our experiment is to compare three elicitation rules: the Quadratic Scoring Rule (QSR), the Matching Probabilities (MP) and the Free Rule (FR). This section is devoted to the presentation of these rules, discussion of their main theoretical properties, and the presentation of their experimental implementation.

3.1 Quadratic Scoring Rule

3.1.1 Definition and properties

In experimental economics, the most commonly used rule is the Quadratic Scoring Rule.[1] We consider here a very simple version of the QSR. Assume a subject reports a confidence level equal to p. She will then win $a - b \times \left(1 - p^2\right)$ if her answer is accurate, and $a - b \times \left(1 - (1-p)^2\right)$ otherwise, where a and b being positive constants.

A scoring rule is said to be *strictly proper* if the unique best strategy of the subject consists in reporting her true subjective probabilities, without any distortion. The QSR is very popular among experimental economists because it is a relatively simple *strictly proper* scoring rule for subjects who use subjective probabilities and maximize their expected reward. It is well known, however, that elicited probabilities through the QSR will be distorted for non-risk neutral subjects. For instance, elicited probabilities for risk averse subjects are expected to be smaller than their subjective probabilities.[2]

3.1.2 Implementation

In our experiment, QSR is implemented as follows. We ask subjects to choose among different levels of remunerations that are presented to them in the Table 1.

Each letter corresponds to a payment scheme (x, y), that yields x if their answer is correct and y if it is not. 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

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

[2] Nevertheless, recent papers try to correct the QSR from risk attitudes (Offerman, Sonnemans, Van de Kuilen, and Wakker (2009), Andersen, Fountain, Harrison, and Rutstrom (2010), Kothiyal, Spinu, and Wakker (2011)).

Choice	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Correct	10	9.98	9.90	9.78	9.60	9.38	9.10	8.78	8.40	7.98	7.5	6.98	6.40	5.78	5.10	4.38	3.60	2.78	1.90	0.98	0
Incorrect	0	0.98	1.90	2.78	3.60	4.38	5.10	5.78	6.40	6.98	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

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 depends on the task. We use cents of euros for the perceptual task, and euros for the quiz.

Note that there is no explicit reference to probabilities in this procedure. Subjects are not told that payment schemes are linked to confidence levels. Moreover, subjective probabilities are not mentioned in the instructions and the QSR theoretical principles are not explained. This is an unusual presentation but we feel it is in line with a revealed preference approach, according to which individual choices among lotteries are the only relevant information. It also avoids a possible drawback of the traditional presentation of the QSR, where options are described in two different terms (payments and reported probabilities), which might induce confusion for the subject. Some experimental evidence support this design. Armantier and Treich (2010) show that using probabilities in the QSR may increase the distortion of elicited probabilities, whereas Offerman, Sonnemans, Van de Kuilen, and Wakker (2009) find no significant difference in probabilities elicited by a QSR with or without explicit reference to probabilities. Finally, observe that three choices (A, K and U, corresponding to probabilities equal to 0, 50 and 100%, respectively) are associated to payments involving two digits numbers, while the other choices involve three digits numbers. One can thus suspect that subjects will concentrate their answers on these choices, because they are simpler. This is certainly true, but previous papers have found a concentration of stated probabilities on 50 and 100%, even with the same number of digits. It is therefore unlikely to be a major issue.

3.2 Matching Probabilities

3.2.1 Definition and properties

The second elicitation rule we consider is the Matching Probabilities, which is a variant of the famous Becker-DeGroot-Marshak mechanism (Becker, Degroot, and

Marschak (1964)). It consists in eliciting an objective probability equivalent to a subjective probability. This principle is known for long (Arrow (1951), Raiffa (1968), Winkler (1972), LaValle (1978) among others ...) but was rarely put in practice until recently (Grether (1992), Abdellaoui, Vossmann, and Weber (2005), Holt (2006), Holt and Smith (2009) are some notable exceptions[3]).

Assume one wants 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;
- a lottery $L(p)$ that gives 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 $L(p)$. Otherwise, the subject is paid according to a lottery $L(q)$ that gives x with probability q and 0 with probability $(1 - q)$.

This procedure provides incentives to truthfully reveal ones' subjective probability. 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	$L(p)$	$L(p)$	$L(q)$
reports $r < p$	$L(p)$	$L(q)$	$L(q)$

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

	$q < p < r$	$p < q < r$	$p < r < q$
reports p	$L(p)$	$L(q)$	$L(q)$
reports $r > p$	$L(p)$	$L(p)$	$L(q)$

One observes 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 Matching Probabilities is that it provides to the subjects incentives to truthfully reveal her subjective probabilities, regardless her attitude towards risk.[4] A drawback, pointed out by Kadane and Winkler (1988), is

[3] Its use has become more widespread in recent years: see among others Dimmock, Kouwenberg, and Wakker (2011), Baillon, Cabantous, and Wakker (2012), Baillon and Bleichrodt (2011), Trautmann and Kuilen (2011), Mobius, Niederle, Niehaus, and Rosenblat (2011).

[4] More details and a formalization can be found in Karni (2009).

that this elicitation rule may not allow to disentangle subjective probabilities from utilities if the agents' wealth is correlated with the event. However, this difficulty cannot arise for the task we consider in our experiment. One main problem is that this rule might seem complicated and thus cognitively demanding. It is then of a particular interest to test whether the complexity of the Matching Probabilities is indeed a problem. As we will see, our data show that such is not the case. Finally, we note that we cannot exclude that subjects might prefer, *ceteris paribus*, to be paid according to their own performance rather than according to an external event. In such a case, they will over-report their subjective probabilities.

3.2.2 Implementation

In practice the Matching Probabilities is implemented using a 0 to 100 scale, with steps of 5 (see Figure 8). After having completed the perceptual task or the quiz, subjects are told that they are entitled with a ticket for a lottery based on their answers' accuracy. In the quiz task, this lottery gives them €10 (€0.10 in the perceptual task) if their answer is correct, and 0 otherwise.

Subjects have then to report on a $0-100$ gauge the minimal percentage of chance p they require to accept an exchange between their lottery ticket and a lottery ticket that gives p chance of winning €10 (€0.10 in the perceptual task). A number l_1 is drawn according to an 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 give them l_1 chance of winning. In this case, a random draw determines the payment: a number l_2 is determined using an uniform distribution between 0 and 100, the lottery is winning if l_1 is higher than l_2. For all trials, the gauge was pre-filled with 75 in order to limit the number of times they have to press the keyboard. The procedure is summarized in Figure 8.

3.3 Free Rule

3.3.1 Definition and properties

The Free Rule just requires the subject to report his confidence, without relating any monetary consequence to stated probabilities. Nothing is done to provide incentives. The main advantage of such a rule is of course its simplicity. It is the less cognitively demanding one, especially comparing to the two previous ones.

The Free Rule is widely used in psychology and neurosciences. In particular, experiments that involve scanning the subjects are very sensitive to response times, as the duration of the experiment is limited and requires a high number of trials to obtain statistically significant results. This makes the Free Rule particularly

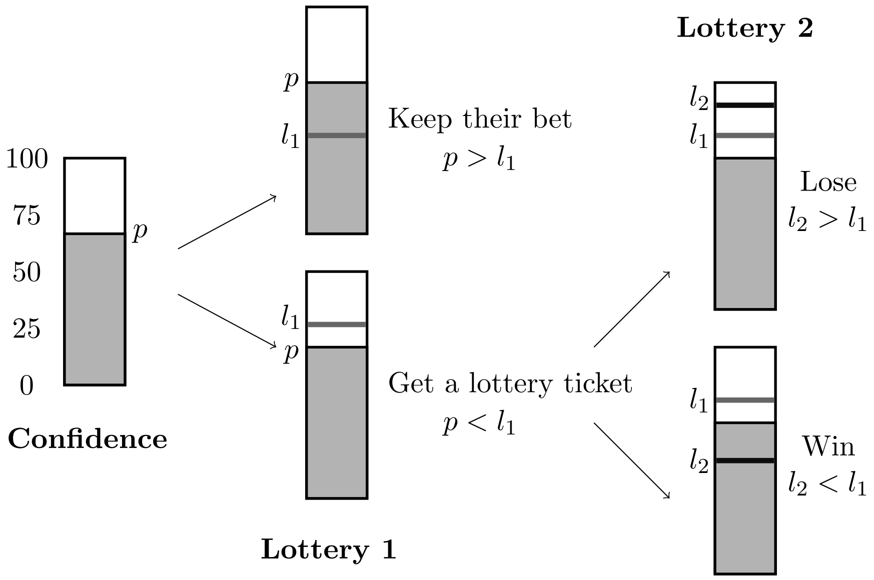

Figure 8: Matching Probabilities

attractive, since probabilities can be elicited very quickly. More generally, there is a trade-off between complexity and incentive compatibility. In contexts where one suspects that incentive compatibility might not be a major issue, it is reasonable to choose a simple rule as the Free Rule. Finally, this rule is also often used in surveys where it is actually difficult to provide monetary incentives.

3.3.2 Implementation

We implement the Free Rule as follows. Subjects just have to choose a level of confidence between 0 and 100 (with steps of 5) on a gauge (see Figure 9). They are told they are free to use the gauge as they want, either by trying to express their confidence level in terms of percentage of chance 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. A correct answer in the quiz provides a payment of €10 (0 if incorrect) and €0.10 in the perceptual task. As for the Matching Probabilities, the gauge was pre-filled with 75 for all trials.

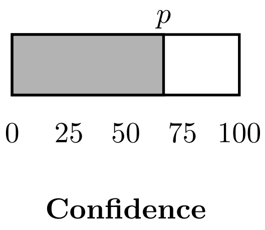

Figure 9: Free Rule

4 Experimental design

4.1 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 has been achieved on computers with 1024×768 screens. We ran two sessions for each rule, that allowed to collect data for 35 to 40 subjects for each rule.

4.2 Stimuli

Our experiment is based on two kinds of task. One is a perceptual task, where subjects are asked to identify which of two circles contains the higher number of dots. The second task is a quiz with questions related to logic and general knowledge. We provide below more details on these tasks.

4.2.1 Perceptual task

The perceptual task we use is a two-alternative forced choice (2AFC) which is known to be a convenient paradigm for SDT analysis (see, e.g., Bogacz, Brown, Moehlis, Holmes, and Cohen (2006)). Subjects have to compare the number of dots contained in two circles (see Figure 1). 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. The position (to the left or the right of the screen) is randomly chosen for each trial. The other one contains $50 \pm \alpha_j$ 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 succeed in 70% of the cases at that level of difficulty. This calibration is done by using a psychophysics staircase (Levitt (1971)). The two other parameters α_1 and α_3 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 if odd.

4.2.2 Quiz task

We elaborated two questionnaires, each containing 30 questions of general knowledge, and 6 logic puzzles. The questions and the puzzles are different in both questionnaire but are similar in terms of difficulty. For instance, one questionnaire contains the question '*Is the distance between London and Tokyo longer than 12000 km?*', whereas the other contains the question '*Is the distance between Paris and Dakar longer than 5000 km?*'. All general knowledge questions have to be answered by "yes" or "no".

4.3 Procedure

In a given experimental session, a single elicitation rule (the same for all subjects) is used. Thus, our study will be based on a between–subjects analysis with a simple 3 × 1 design.[5]

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 by giving their choice and their confidence in this choice. They have no feedback on their answers.

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

[5] Pilot experiments have shown that subjects get confused if one ask them to use different elicitation rules.

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. Observe that this experimental design allows to investigate learning effects. For instance, we can examine whether subjects' metacognitive abilities are higher in the second quiz compared to the first quiz.

4.4 Payment

The payment contains three parts. There is a show-up fee of €5. The quiz tasks are paid as follows. One question is randomly selected at the end of the experiment, and payments are based on the answer given to that question. Such a procedure is standard, and allows to avoid edging strategies. For groups using the QSR or the Matching Probabilities, payments are computed according to the elicitation rule used for the selected question, with a maximum payment of €10 and a minimum of €0. Subjects in the group using the Free Rule are paid €10 if their answer to the selected question is correct, and €0 otherwise. For the perceptual task, subjects are paid for each trials. For groups using the QSR or the Matching probabilities, each 100 trials 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 Free Rule are paid 0.10 € for each correct answer.

5 Results

We drop the results for 6 subjects out of 113 because their stated confidence did not vary: 3 in the QSR group, 2 in the Free Rule group and 1 in the Matching Probabilities group. The three groups are similar according to demographic data. There is no significant statistical difference between groups in the mean success rate for the perceptive and the quiz tasks.

The results are presented in three parts. First, as a preliminary step, we perform some descriptive analysis to draw a general picture of elicited probabilities. Second, we investigate how elicitation rules perform with respect to predictions related to SD-confidence. We conclude our analysis with some further results concerning cross-tasks comparisons.

5.1 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 FR and the MP are similar, the one corresponding to the QSR differs significantly (see Figure 10). 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 third of elicited probabilities are either equal to 50% or 100% when one uses the QSR, which is twice as much as for the two other rules. One can also observe that the FR yields to a greater concentration on 75%. This is likely to be explained by the fact that the gauge was pre-filled precisely at this value. However, we do not observe such a result for the MP, that is also based on a jauge pre-filled at 75%. We suspect that this is due to the fact that no incentive is provided in the FR, and that this might lead subjects to simply not make the effort to change this value in many cases.

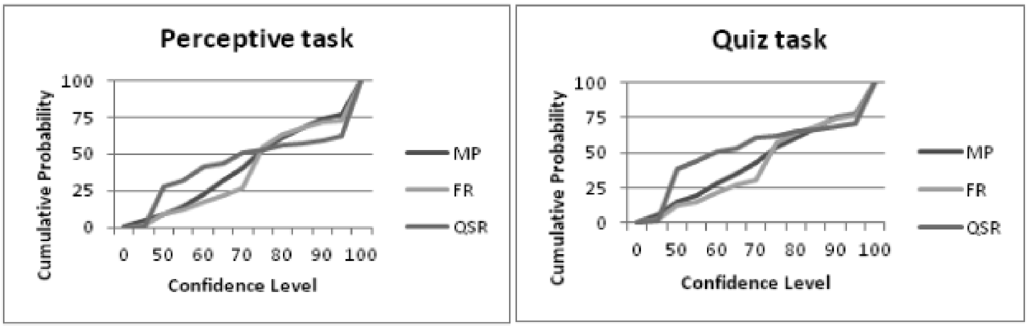

Figure 10: Cumulative probability distribution of elicited confidence.

Let us next have a look at how subjects' stated confidence is related to their actual success rate (see Figure 11). A first observation is that, whatever the elicitation rule used, subjects are globally overconfident. Moreover, the difference between stated confidence and observed success rates increases with stated confidence. If we consider all the 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 slightly higher than 50%. Finally, we note that none of the elicitation rules provides strictly increasing relationship between stated confidence and actual success rate.

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. One may thus investigate whether the results described above are robust to learning effects.

During the second part of the experiment, subjects used 100 times the elicitation rule with feedback. They could thus have learn to use the elicitation rule during this part. We can therefore measure learning effects by comparing (*i*) results for the quiz in the first and last parts of the experiment, and (*ii*) results for the first half

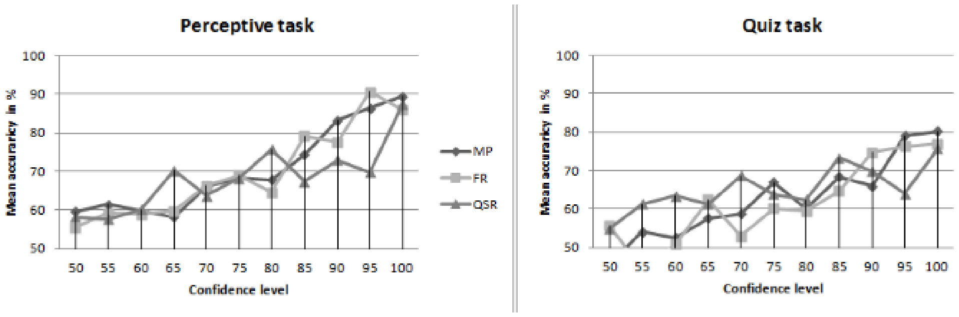

Figure 11: Stated confidence vs. accuracy

This figure represents the mean accuracy for each level of confidence between 50% and 100% with step of 5.

(first 50 trials) and the second half (50 last trials) of the perceptive task. Table 2 provides details about the learning effect for the three different rules.

Rule	AU2ROC quiz	AU2ROC perceptual task
MP_Part1	0.6351 (.0788)	0.6677 (.0915)
MP_Part2	0.6548 (.0981)	0.6792 (.0864)
MP_(Part2 - Part1)	0.0197 (0.1715)	0.0114 (0.1918)
QSR_Part1	0.6080 (.1046)	0.6647 (.0767)
QSR_Part2	0.6272 (.0863)	0.6860 (.0873)
QSR_(Part2 - Part1)	0.0193 (0.2251)	0.0212 (0.1011)
FR_Part1	0.6382 (.1157)	0.6570 (.1005)
FR_Part2	0.6092 (.1058)	0.6769 (.0929)
FR_(Part2 - Part1)	-0.0290 (0.1249)	0.0199 (0.1247)
AllRule_(Part2 - Part1)	0.0036 (.3943)	0.0172 (.0263)**

Table 2: Learning effect: AU2ROC for quiz and perceptual task

We observe some evidence of learning effects for discrimination ability. This is the case for the three rules in the perceptual task. Since the increase is 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. A similar increase is observed in the quiz task for the MP and the QSR but not for the FR. This lack of improvement of metacognitive ability might be due to boredom to continue to report confidence level without incentive in a repetitive task. Overall, we thus conclude that there is no difference in terms of learning between the MP and the QSR.

5.2 SD-confidence

We now consider whether elicitation rules yield individuals to report confidence levels that are compatible with the predictions of SDT. In other words, we investigate to what extent we could interpret the confidence levels reported by individuals as related to the probabilities they use in order to perform the perceptive task.

The first thing we need is to compute predicted SD-confidence in the perceptive task. The only difficulty here is that there are actually five difficulty levels. We extend the bayesian analysis described in section 2 to this case, under the assumption that subjects have correct priors on the distribution of difficulty levels. We can examine now how good are the three rules into eliciting SD-confidence. We proceed by examining in turn each of the predictions A1 to A5.

Let us start with prediction A1, which sates that elicited confidence should be close to predicted SD-confidence. A first answer is given by comparing elicited confidence and predicted SD-confidence distributions. We report in Figure 12 the elicited confidence and predicted SD-confidence distributions for each elicitation rule (date are pooled across all levels of difficulty and all subjects). It appears clearly that the MP is the rule that yields to the best fit. The FR is plagued by the large proportion of elicited confidence levels equal to 75%, which is the pre-filled value of the gauge. 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 observe also a high peak at the 100% value (with 38% of the answers), which cannot be explained by risk aversion, and do not correspond to predictions of SDT (only 18% of the answer should take this value according to SDT).

To confirm the visual feeling that MP yields 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 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 Table 3. The results for t-tests (not displayed in the table) show that the two distances are significantly lower (at a level of 1% for the Chi-Square distance, and 5% for the Kolmogorov-Smirnov distance) for the MP data than for the QSR and the FR data, while there are no significant difference between QSR and FR data. We also found that the two distances are strongly correlated (corr = 0.85).

The second prediction (Prediction A2) states that elicited AU2ROC should be close to predicted ones. We display in Figure 13 the corresponding data for each rule.

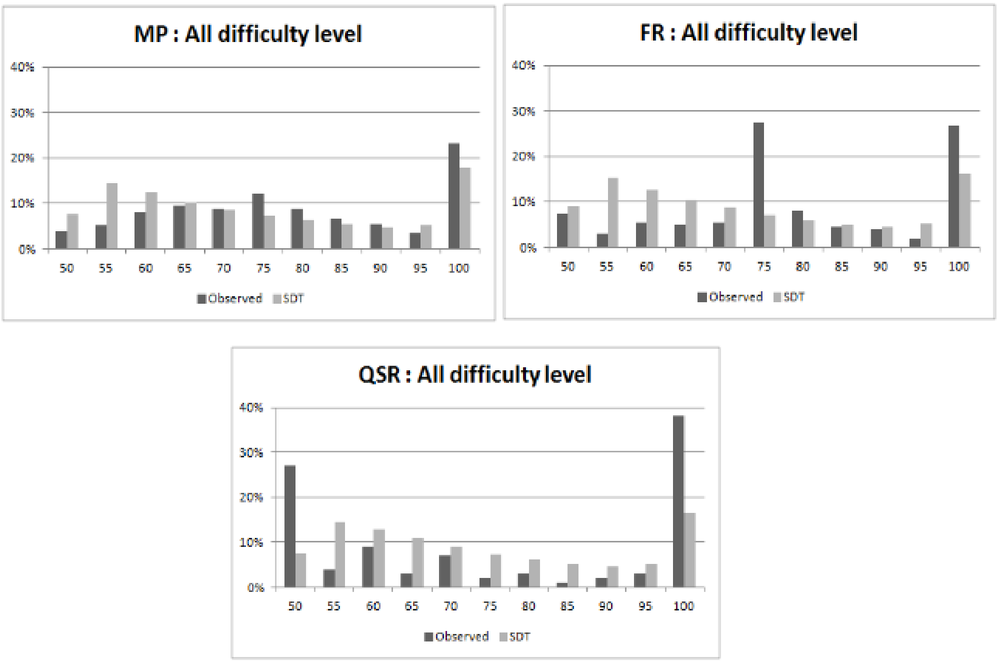

Figure 12: Elicited confidence and predicted SD-confidence distributions

Distance between predicted and observed confidence distances	MP (n = 39)	QSR (n = 33)	FR (n = 35)
χ^2square distance	0.48(.31)	0.81(.37)	0.76(.49)
Kolmogorov-Smirnov distance	0.32(.14)	0.39(.16)	0.40(.22)

Table 3: Distances between elicited confidence and predicted SD-confidence distributions

The correlation between observed and predicted AU2ROC is positive and statistically significant for the MP (corr = 0.39) and for the FR (corr = 0.40) while it is negative but not statistically significant for the QSR data (corr = - 0.11).

Our third prediction (A3) is that observed AU2ROC should not be greater than the predicted one. This is actually the case for 33 out of 39 subjects (85%) in the MP group, 26 out of 35 (74%) in the FR group and 24 out of 33 (73%) in the QSR group.[6]

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 (Prediction A4). In other words,

[6] The difference between MP and QSR is statistically significant at 20%.

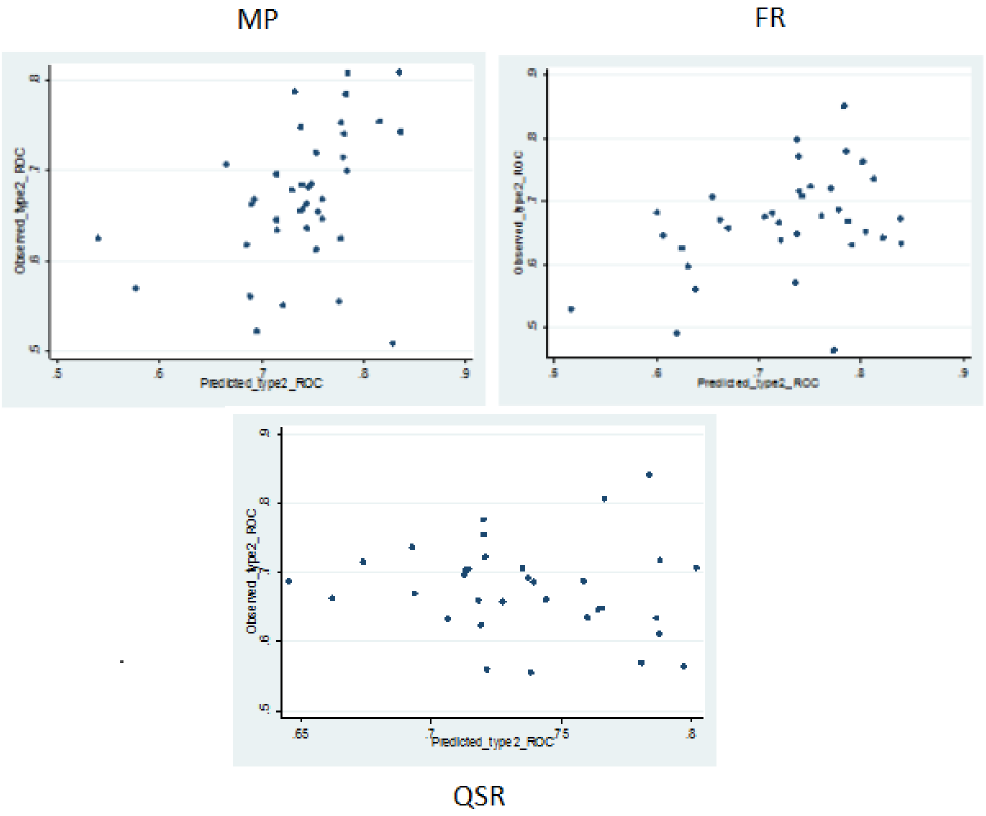

Figure 13: Area under observed and predicted type 2 ROC

we should observe a positive correlation between the distance between observed and predicted confidence distributions on 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:

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

We report the correlations in Table 4. We observe a positive and significant correlation for both the MP and the FR. On the other hand, the results are less conclusive

Corr. btw. *ROC_dist* and...	MP (39)	QSR (33)	FR (35)
... χ^2square distance	0.42***	-0.08	0.50***
... Kolmogorov-Smirnov distance	0.47***	0.30*	0.34**

Table 4: Correlations between χ^2 distance or Kolmogorov-Smirnov distance and ROC distance

*** (resp. **,*) means a level of significance at 1% (resp. 5%,10%).

for the QSR, for which we observe no correlation between distances measured by the χ^2 metric.

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 (Prediction A5). We report these correlations in Table 5. We found that performances in type 1 and type 2 tasks are strongly correlated when confidence is elicited with MP. The correlation is still positive, but less significant for the FR. But we found no correlation between performances in type 1 and type 2 tasks when the QSR is used.

MP (39)	QSR (33)	FR (35)
0.41***	-0.10	0.32*

Table 5: Correlations between mean success rate and observed AU2ROC
*** (resp. *) means a level of significance at 1% (resp. 10%).

Taken together, our results suggest that elicitation rules strongly differ in the kind of confidence they allow to report. Whereas confidence levels reported using MP are globally compatible with predicted SD-confidence, those obtained through QSR can hardly be explained by SDT. The results concerning the FR 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.

5.3 Extension to non perceptive task

An important question is whether the models and results described above generalize to other tasks. We have no simple answer to that question. However, we note that SDT has been routinely and successfully used in experimental psychology for non-perceptive task, such as memory or recognition task. Moreover, it has been recently shown that individual metacognitive performances in different perceptive tasks are correlated, suggesting that there exist general processes related to metacognition, and independent on the specific task considered (Song, Kanai, Fleming, Weil, Schwarzkopf, and Rees (2011)). Of course, this evidence only applies to perceptive tasks, and it is to the best of our knowledge an open question whether it also holds for non-perceptive task.

This leads us to make here the assumption that our analysis also applies to non-perceptive task. Because our subjects perform both a perceptive and a non-perceptive task, a validation of this assumption (if any) would be found in our data. We briefly look at how performances of elicitation rules are correlated across tasks. The main correlations are reported in Table 6

Correlation between	MP (39)	QSR (33)	FR (35)
AU2ROC in quiz and perceptive task	0.44***	0.10	0.49***

Table 6: Correlations between AU2ROC in quiz and perceptive task.
*** means a level of significance at 1%.

We observe a strong and positive correlation between AU2ROC in the quiz and in the perceptive task for both the MP and the FR. As we have observed that MP (and to a lesser extent, FR) are reasonably good in eliciting subjects' SD-confidence in the perceptive task, we take this result as a piece of evidence that our SDT analysis can actually be extended to other, less simple, tasks, as the quiz task. It is not surprising that we do not observe such a positive correlation for the QSR, as we have shown that it does poorly in terms of elicitation of SD-confidence.

6 Conclusion

Signal Detection Theory provides a theoretical model for predicting individuals confidence (expressed as probabilities) from the observation of their performances in a simple perceptive task. This model is moreover consistent with behavioral and neuronal evidence. A natural question is whether one can actually elicit subjective probabilities that are close to those predicted by SDT.

We elicited individuals' confidence in a simple perceptive task using three different rules: the Quadratic Scoring Rule, the Matching Probabilities, and the Free Rule. We found that MP provided results remarkably close to those predicted by SDT. We also observed that although this rule might seem complicated at first sight, there is no evidence that subjects had more difficulties using MP than the two other rules. Moreover, using a task based on a knowledge quiz for the same subjects, we found evidence suggesting that these results might possibly extend to other, non perceptive, tasks. MP thus appears as a good candidate if one seeks for an incentive-compatible elicitation rule that is not affected by rewards, and remains reasonably simple to use.

In this paper, we focused on the elicitation of probabilities actually used by individuals in their decision processes. We should note that one could ask a different question, and seek for a rule that leads subjects to provide a good evaluation of their own performance in the perceptive or quiz task. The ability to be a good assessor of one's own performance is known as calibration and discrimination. A precise answer to this question would require to be able to compare individuals performances for different elicitation rules, i.e., to have intra-individual data. It is thus beyond the scope of this paper, since we do not have such data.[7] While we keep this question for future research, a first rough answer can be given by simply comparing AU2ROC for the three rules. The results are reported below in Table 7. We only found a statistically significant difference for the quiz task. The MP performs better than the QSR. This result is not very conclusive, as we found no significant difference between FR and MP, nor between QSR and FR.

Rule	AU2ROC for quiz taske	AU2ROC for perceptive task
MP (n=39)	0.6439 (.0623)	0.6702 (.0754)
QSR (n=33)	0.6148 (.0615)	0.6769 (.0664)
FR (n=35)	0.6233 (.0881)	0.6674 (.0812)
(MP - QSR)	0.0291 (0.0509)*	-0.0067 (.6944)
(MP - FR)	0.0206 (0.2465)	0.0029 (.8751)
(QSR - FR)	-0.0085 (0.6462)	0.0095 (.5991)

Table 7: Rules comparison: discrimination

This table provides the mean AU2ROC for the three rules (with s.d.). The rules are compared by pairs with a test of difference (t-test with the p-value in parenthesis). * means a level of significance at 10%.

We complete these results with another well-known measure of the quality of elicitation rules, namely the calibration index. Consider a subject who stated subjective probabilities about n events, p_i being her stated probability for event E_i, e_i being the indicator variable that takes value 1 if she accurately predicts event E_i.

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

We reported calibration for quiz and perceptive task in Table 8.

We observe that the QSR performs better in terms of calibration for the quiz: it displays a lower degree of overconfidence than the two other rules. However, the result is not very conclusive, as it does not hold for the perceptive task.

Regressions of the calibration index on a set of explanatory variables that includes dummies for the elicitation rules, individual mean performances in the different tasks (knowledge, logic, perception) and demographic variables yield to the same result. Therefore, on the basis of the available evidence, we cannot conclude that one rule clearly dominates the others in terms of calibration quality. Comparing the rules in that respect will thus require additional data. We will investigate this

[7] In pilot experiments, we found that subjects got confused if one asks them to use different elicitation rules during the same session. Obtaining intra-individual data thus requires to have the same individuals attending several sessions, separated by a sufficient laps of time. It thus implies a specific design.

Rule	Calibration index for quiz task	Calibration index for perceptive task
MP (n=39)	0.1113 (.0886)	0.0423 (.0889)
QSR (n=33)	0.0612 (.1284)	0.0417 (.0932)
FR (n=35)	0.1372 (.1092)	0.0735 (.1104)
(MP - QSR)	0.0501 (0.0552)*	0.0006 (0.9785)
(MP - FR)	-0.0259 (0.2647)	-0.0312 (0.1822)
(QSR - FR)	-0.0760 (0.0105)**	-0.0317 (0.2063)

Table 8: Rules comparison: calibration

This table provides the mean calibration index for the three rules (with s.d.). The rules are compared by pairs with a test of difference (t-test with the p-value in parenthesis). ** and * mean respectively a level of significance at 5% and 10%.

question in future experiments.

References

Abdellaoui, M., F. Vossmann, and M. Weber (2005): "Choice-Based Elicitation and Decomposition of Decision Weights for Gains and Losses under Uncertainty," *Management Science*, 51(9), 1384–1399.

Andersen, S., J. Fountain, G. Harrison, and E. Rutstrom (2010): "Estimating Subjective Probabilities," *CEAR Working Paper*.

Armantier, O., and N. Treich (2010): "Eliciting Beliefs: Proper Scoring Rules, Incentives, Stakes and Hedging," *Working Paper*.

Arrow, K. J. (1951): "Alternative Approaches to the Theory of Choice in Risk-Taking Situations," *Econometrica*, 19, 404–437.

Baillon, A., and H. Bleichrodt (2011): "Testing Ambiguity Models through the Measurement of Probabilities for Gains and Losses," *Working Paper*.

Baillon, A., L. Cabantous, and P. Wakker (2012): "Aggregating Imprecise or Conflicting Beliefs: An Experimental Investigation Using Modern Ambiguity Theories," *Journal of Risk and Uncertainty (in press)*.

Beck, J., W. Ma, R. Kiani, T. Hanks, A. Churchland, J. Roitman, M. Shadlen, P. Latham, and A. Pouget (2008): "Probabilistic population codes for Bayesian decision making," *Neuron*, 60(6), 1142–1152.

Becker, G. M., M. H. Degroot, and J. Marschak (1964): "Measuring utility by a single-response sequential method," *Behavioral Science*, 9(3), 226–232.

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

Brainard, D. (1997): "The Psychophysics Toolbox," *Spatial Vision*, 10, 433–436.

Clarke, F., T. Birdsall, and W. Tanner (1959): "Two types of ROC curves and definitions of parameters," *The Journal of the Acoustical Society of America*, 31(5), 629–630.

Dimmock, S., R. Kouwenberg, and P. Wakker (2011): "Ambiguity Attitudes and Portfolio Choice: Evidence from a Large Representative Survey," *Netspar Discussion Paper No. 06/2011-054*.

Fleming, S., and R. Dolan (2010): "Effects of loss aversion on post-decision wagering: Implications for measures of awareness," *Consciousness and cognition*, 19(1), 352–363.

Fleming, S., J. Huijgen, and R. Dolan (2012): "Prefrontal Contributions to Metacognition in Perceptual Decision Making," *The Journal of Neuroscience*, 32(18), 6117–6125.

Fleming, S., R. Weil, Z. Nagy, R. Dolan, and G. Rees (2010): "Relating introspective accuracy to individual differences in brain structure," *Science*, 329(5998), 1541–1543.

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

Gold, J., and M. Shadlen (2002): "Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward," *Neuron*, 36(2), 299–308.

Green, D., and J. Swets (1966): *Signal detection theory and psychophysics*, vol. 1974. Wiley New York.

Grether, D. (1992): "Testing Bayes rule and the representativeness heuristic: Some experimental evidence," *Journal of Economic Behavior and Organization*, 17, 31–57.

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.

Kadane, J. B., and R. L. Winkler (1988): "Separating Probability Elicitation From Utilities," *Journal of the American Statistical Association*, 83(402), 357–363.

Kanai, R., V. Walsh, and C. Tseng (2010): "Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness," *Consciousness and cognition*, 19(4), 1045–1057.

Karni, E. (2009): "A Mechanism for eliciting Probabilities," *Econometrica*, 77(2), 603–606.

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

Kiani, R., and M. Shadlen (2009): "Representation of confidence associated with a decision by neurons in the parietal cortex," *Science*, 324(5928), 759–764.

Kothiyal, A., V. Spinu, and P. Wakker (2011): "Comonotonic Proper Scoring Rules to Measure Ambiguity and Subjective Beliefs," *Journal of Multi-Criteria Decision Analysis*, 17, 101–113.

LaValle, I. H. (1978): *Fundamentals of Decision Analysis*. Holt, Rinehart and Winston, New York.

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

Ma, W., J. Beck, P. Latham, and A. Pouget (2006): "Bayesian inference with probabilistic population codes," *Nature neuroscience*, 9(11), 1432–1438.

Maniscalco, B., and H. Lau (2012): "A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings.," *Consciousness and cognition*, 21(1), 422.

Mobius, M., M. Niederle, P. Niehaus, and T. Rosenblat (2011): "Managing Self-Confidence: Theory and Experimental Evidence," *NBER Working Paper No. 17014*.

Nieder, A., and S. Dehaene (2009): "Representation of number in the brain," *Annual review of neuroscience*, 32, 185–208.

Nieder, A., D. Freedman, and E. Miller (2002): "Representation of the quantity of visual items in the primate prefrontal cortex," *Science*, 297(5587), 1708–1711.

Nieder, A., and E. Miller (2003): "Coding of Cognitive Magnitude:: Compressed Scaling of Numerical Information in the Primate Prefrontal Cortex," *Neuron*, 37(1), 149–157.

——— (2004): "A parieto-frontal network for visual numerical information in the monkey," *Proceedings of the National Academy of Sciences of the United States of America*, 101(19), 7457.

Nyarko, Y., and A. Schotter (2002): "An Experimental Study of Belief Learning Using Elicited Beliefs," *Econometrica*, 70(3), 971–1005.

Offerman, T., J. Sonnemans, G. Van de Kuilen, and P. Wakker (2009): "A Truth-Serum for Non-Bayesian: Correcting Proper Scoring Rules for Risk Attitudes," *Review of Economic Studies*, 76(4), 1461–1489.

Palfrey, T., and S. Wang (2009): "On Eliciting Beliefs in Strategic Games," *Journal of Economic Behavior and Organization*, 71(2), 98–109.

Piazza, M., V. Izard, P. Pinel, D. Le Bihan, and S. Dehaene (2004): "Tuning curves for approximate numerosity in the human intraparietal sulcus," *Neuron*, 44(3), 547–555.

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

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

Raiffa, H. (1968): *Decision Analysis.* Addison-Wesley, London.

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

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

Trautmann, S., and G. v. d. Kuilen (2011): "Belief Elicitation: A Horse Race among Truth Serums," *Working Paper.*

Winkler, R. L. (1972): *An Introduction to Bayesian Inference and Decision Theory*. Holt, Rinehart and Winston, New York.

Yang, T., and M. Shadlen (2007): “Probabilistic reasoning by neurons,” *Nature*, 447(7148), 1075–1080.