A short-but-efficient test for overconfidence and prospect theory. Experimental validation

Peon, David and Calvo, Anxo and Antelo, Manel

University of A Coruna, University of A Coruna, University of Santiago de Compostela

5 March 2014

A short-but-efficient test for overconfidence and prospect theory. Experimental validation[*]

David Peón[ab], Anxo Calvo[c], Manel Antelo[d]

This version: February 2014

Abstract

Two relevant areas in the behaviorist literature are prospect theory and overconfidence. Many tests are available to elicit their different manifestations: utility curvature, probability weighting and loss aversion in PT; overestimation, overplacement and overprecision as measures of overconfidence. Those tests are suitable to deal with single manifestations but often unfeasible, in terms of time to be performed, to determine a complete psychological profile of a given respondent. This paper contributes to provide two short tests, based on classic works in the literature, to derive a complete profile on prospect theory and overconfidence.

We conduct an experimental research with 126 students to validate the tests, revealing they are broadly efficient to replicate the regular results in the literature. The experimental analysis of all measures of overconfidence and prospect theory using the same sample of respondents allows us to provide new insights on the relationship between these two areas. Finally, enhancements for future research are suggested.

Keywords: Experimental economics, overconfidence, prospect theory, behavioral finance, utility measurement, overestimation, overplacement, overprecision

JEL Classification: D03, D81, C91

[*] The authors want to thank Paulino Martínez and Xosé Manuel M. Filgueira for their very valuable support in the experiment design. Authors also want to thank Juan Vilar and Jose A. Vilar for technical assistance.

[a] Grupo BBVA and Departamento de Economía Financeira e Contabilidade, Universidade da Coruña, Campus Elviña s/n, 15071 A Coruña. Email: *david.peon@udc.es*

[b] Corresponding author.

[c] Departamento de Economía Financeira e Contabilidade, Universidade da Coruña, Campus Elviña s/n, 15071 A Coruña. Email: *anxo.calvo@udc.es*

[d] Department of Economics, University of Santiago de Compostela, Campus Norte, 15782 Santiago de Compostela (Spain). E-mail: *manel.antelo@usc.es*

1. INTRODUCTION

Behavioral biases have been suggested to explain a wide range of market anomalies. A recent and growing field is the analysis of overconfidence effects on credit cycles (e.g., Rötheli, 2012). An interesting step forward would be to obtain experimental evidence of whether behavioral biases by participants in the banking industry could feed a risk-seeking behavior that explains, up to some extent, the excessive lending by retail banks.

To that purpose, we organized a series of experimental sessions that were divided in two parts. The first part was a set of questions devised to determine the psychological profile, based on prospect theory and overconfidence, of each participant. The second part was a strategy game designed to replicate in an experimental setting the basics of the decision-making process of a bank that grants credit to costumers under conditions of risk and uncertainty. Results of the second part are analyzed elsewhere (Peón et al., 2014).

The main motivation of this paper is to design, for the first part of the experiment, some simple tests on overconfidence and prospect theory. We base our work in this paper on some classics in the literature. First, regarding prospect theory we follow Tversky and Kahneman (1992) –who develop cumulative prospect theory, CPT- and Abdellaoui et al. (2008) –who provide an efficient method to measure utility under CPT. Second, for overconfidence we follow Moore and Healy (2008) –who identify three measures widely accepted since then- and Soll and Klayman (2004) –who provide a method to disentangle variability and true overconfidence in interval estimates.

However, trying to replicate the original tests in the experiment would be unfeasible. To illustrate, Tversky and Kahneman (1992) mention that subjects in their experiment "participated in three separate one-hour sessions that were several days apart" (p. 305) in order to complete a set of 64 prospects, while participants in the experimental test by Moore and Healy (2008) spent "about 90 minutes in the laboratory" to complete 18 rounds of 10-item trivia quizzes. Consequently, we need shorter versions of these tests, in a way the number of items required for estimation purposes are reduced but they do not compromise efficient results. Indeed, the concern to design tests that are shorter and more efficient is a classic in the behaviorist literature (e.g., Abdellaoui et al., 2008), since they would enhance the scope for application of behavioral theories.

This paper is devoted to explain how shorter tests were designed to obtain a basic profile, in terms of CPT and overconfidence, of a given individual, and the literature that supports our choices. Furthermore, the tests were implemented to a sample of 126 under- and postgraduate students in the University of A Coruña (UDC) during October 2013. The experiment results will be determinant to assess the quality of data we obtained by comparing them with regular results in the literature.

Three main contributions of this paper are in order. First, we design two short tests on overconfidence and CPT that are able to elicit the three measures of overconfidence (overestimation, overplacement and overprecision) as well as the complete set of parameters in prospect theory –namely, utility curvature, probability weighting and loss aversion. Second, we conduct an experimental research with 126 students to validate the tests. In the bulk of this paper, we conduct a simplicity – efficiency tradeoff analysis by comparing our results with those regular in the literature. Third, the experimental analysis of all measures of overconfidence and prospect theory using the same sample is something that, to the best of our knowledge, was not done before. This allows us to provide new insight on the relationship between these two relevant areas in the behavioral literature.

The structure of the article is as follows. In Section 2, after briefly introducing theory and current state of the art, we describe how our tests were designed, first on overconfidence and then on prospect theory. In Section 3 we discuss the experiment results and the reliability of the tests designed according to experimental evidence. Section 4 tests some hypothesis about the relationship between demographic priors and behavioral variables. Section 5 concludes.

2. OVERCONFIDENCE AND PROSPECT THEORY: THEORY AND EXPERIMENT DESIGN

2.1. Overconfidence

The prevalence of overconfidence among people is a classic in the behavioral literature. Moore and Healy (2008) identify three different measures of how people may exhibit overconfidence: in estimating their own performance (*overestimation*); in estimating their own performance relative to others (*overplacement* or 'better-than-average' effect); and having an excessive precision to estimate future uncertainty (*overprecision*).

For test design, we approach to overconfidence following Moore and Healy (2008)'s theory for several reasons. First, the clarification of the three measures overconfidence has been widely accepted since then. Second, they were able to make a synthesis of the previous debate between ecological and error models versus the cognitive bias interpretation, offering a model that applies the Bayesian principle of updating beliefs from prior beliefs based on data. Third, their model is able to predict both over- and underconfidence in two of their different manifestations (estimation and placement) as well as the hard-easy effect. Finally, their tests are really simple, allowing us to implement a highly efficient test that requires only a few minutes to perform it.

Our tests are also designed taking into consideration some consensus in the literature regarding two aspects. First, frequency judgments across a set of items are less prone to overconfidence than are judgments of correctness at the item-level (where participants are required to provide a probabilistic judgment). Second, the hard-easy effect: on easy tasks, people underestimate their performance but overplace themselves compared to others; hard tasks, instead, produce overestimation and underplacement. In order to account for these two discussions, Moore and Healy (2008) conduct their tests asking for frequency judgments across several sets of items of easy, medium and hard difficulty.

Overprecision, on the other hand, requires an alternative analysis. A classic approach is to ask for interval estimates (Soll and Klayman, 2004), as opposed to binary choices. Using binary choices causes overestimation and overprecision to be “one and the same” (Moore and Healy, 2008), because being excessively sure you got the correct answer from a choice of two reflects both overestimation of your performance and excessive confidence in the precision of your knowledge. Consequently, in our tests, in order to avoid confusing overestimation and overprecision, we study overestimation by measuring perceptions across a set of items, whereas overprecision is analyzed through a series of questions on interval estimates.

Test design

The tests devoted to elicit the overconfidence factors will consist of a set of trivial-like questions, devised to determine the degree of overestimation, **E**, and overplacement, **P**, of each respondent, plus a set of additional questions where subjects are asked to provide some confidence interval estimations –devised to determine the degree of overprecision **M** (following notation by Soll and Klayman, 2004) of each respondent.

Our tests for **E** and **P** are a simplified version of Moore and Healy (2008)’s trivia tests. Indeed, several questions have been taken from the original tests by the authors.[1] In order to elicit the parameters **E** and **P** of each respondent, participants are required to complete

[1] We would like to thank the authors for providing their tests online, they have been really helpful to us. We would like to be equally helpful to other researchers: the complete set of questions in our tests will be freely available at www.dpeon.com/documentos

a set of 4 trivial-like games with 10 items each one. In order to account for the hard-easy effect, 2 quizzes were of easy difficulty and 2 of hard difficulty –though obviously this information should not be provided to participants. Since answers to questions involving general knowledge tend to produce overconfidence, while responses to perceptual tasks often result in underconfidence (Stankov et al., 2012), we asked questions of general knowledge under a time-constrained situation (a time limit of 150 seconds per trivia) to have a somehow mixed scenario.

Prior to solving the trivia, participants were instructed and asked to answer a practice question to familiarize with the experimental setting. Then they took the actual quizzes. When time was over, they were required to estimate their own scores, as well as the score of a *randomly selected previous participant* (RSPP).[2] Finally, they repeated the same process for all the other three rounds.

Overestimation is calculated by substracting a participant's actual score in each of the 4 trivia from his or her reported expected score, namely

$$\mathbf{E} = E[X_i] - x_i \tag{1}$$

where $E[X_i]$ is individual i's belief about his or her expected performance in a particular trivia test, and x_i measures his or her actual score in that test. We calculate (1) for each of the 4 trivia, and then sum all 4 results. A measure **E**>0 means the respondent exhibits overestimation, while **E**<0 means underestimation. Additional information on the hard-easy effect may be available if similar estimations are calculated separately for the hard and easy tasks, in order to see if **E** is negative on easy tasks and positive on hard ones.

Overplacement is calculated taking into account whether a participant is really better than others. For each quiz we use the formula

$$\mathbf{P} = (E[X_i] - E[X_j]) - (x_i - x_j) \tag{2}$$

where $E[X_j]$ is that person's belief about the expected performance of the RSPP on that quiz, and x_j measure the actual scores of the RSPP. We calculate (2) for each of the 4 trivia, and then sum all 4 results. A measure **P**>0 means the respondent exhibits overplacement, while **P**<0 means underplacement. Again, additional information on the hard-easy effect may be available if similar estimations are calculated separately for the hard and easy tasks, in order to see if **P** is positive on easy tasks and negative on hard ones.

[2] They were required to estimate 'the average score of other students here today and in similar experiments with students of this University'.

Finally, overprecision is analyzed through a separate set of 6 questions devised following Soll and Klayman (2004). Tests for overprecision usually require participants to provide confidence intervals around the subjects' answers. However, Soll and Klayman (2004) show overconfidence in interval estimates may result from variability in setting interval widths. Consequently, in order to disentangle variability and true overprecision, they define the ratio

$$\mathbf{M} = \text{MEAD/MAD}, \tag{3}$$

to estimate overprecision, where MEAD is the mean of the expected absolute deviations implied by each pair of fractiles a subject gives, and MAD the observed mean absolute deviation. Thus, **M** represents the ratio of observed average interval width to the well-calibrated zero-variability interval width. Consequently, **M** = 1 implies perfect calibration, and **M**<1 indicates an overconfidence bias that cannot be attributed to random error, with the higher overprecision the lower *M* is.

Soll and Klayman show that different domains of questions are systematically associated with different degrees of overconfidence (which highlights the risks of relying on any single domain) and that asking subjects for three fractile estimates (two boundaries and a median estimate) rather than two reduces overconfidence. With these results in mind, we designed our test as follows. First, in each question we ask participants to specify a three-point estimate (median, 10% and 90% fractiles, so we have low and high boundaries for an 80% confidence interval). Second, Soll and Klayman ask a set of several questions per domain. However, since we can only ask a few questions and the risks of relying on a single domain were emphasized, we choose to make only a pair of questions on three different domains. This causes a problem regarding the statistical reliability of *M* that will be discussed below.

Questions 1 to 4 are traditional almanac questions –i.e., general knowledge questions on arbitrarily chosen topics- on two different domains. The first domain replicates two questions by Soll and Klayman (2004) about 'the year in which a device was invented'. The second one asked about mortality rates –a classic question about shark attacks (Shefrin, 2008) plus another one regarding road accidents in Spain. Questions 5 and 6 try an alternative approach. Most studies of confidence ask judges to draw information only from their knowledge and memory. Soll and Klayman introduce a variation: including domains for which participants could draw on direct, personal experience. We do the same to ask, again inspired by Soll and Klayman, about 'time required to walk from one place to another in A Coruña at a moderate (5 km/h) rate without interruption'. Participants were required in all six cases to provide a median estimate and an 80% confidence interval around their answers.

The procedure we implement to estimate *M* is as follows. We use a beta function to estimate the implicit subjective probability density function, SPDF, of each respondent. Then we estimate MEAD and MAD. First, for each question we calculate the expected surprise implied by the SPDF to obtain the expected absolute deviation, *EAD*, from the median. Then, the mean of the *EAD*s for all questions in a domain is calculated, *MEAD*. Second, for each question we calculate the observed absolute deviation between the median and the true answer, and then the mean absolute deviation, *MAD*, of all questions in a same domain. Then we calculate the ratio **M** for each domain. Consequently, we have 3 different estimations of the ratio. *M* could then simply be calculated as either the average ($\mathbf{M}_{\mathbf{avg}}$) or the median ($\mathbf{M}_{\mathbf{med}}$) of the 3 different estimations.

2.2. Prospect theory

Prospect theory is the best known descriptive decision theory. Kahneman and Tversky (1979) provide extensive evidence that, when making decisions in a context of risk or uncertainty, most individuals (i) show preferences that depend on gains and losses with respect to a reference point, and (ii) form beliefs that do not correspond to the statistical probabilities (their perception of the risks associated with a decision may be biased).

Assume two mutual exclusive states of the world, s_1 and s_2 (state s_1 occurring with probability *p*, $0<p<1$) and consider a simple binary lottery with payoff c_1 in s_1 and c_2 in s_2, $c_1< c_2$. PT changes both the way utility is measured –providing a value function v(·) that is defined over changes in wealth- and the way subjects perceive the probabilities of the different outcomes – by applying a probability weighting function, *w(p),* to the objective probabilities *p*, as follows

$$w(p)\cdot v(c_1) + w(1\text{-}p)\cdot v(c_2) \qquad (4)$$

On one hand, the value function has three essential characteristics (reference dependence, diminishing sensitivity and loss aversion) that result in the well-known shape that is kinked at the reference point, concave above, convex below, and steeper in the negative domain. On the other hand, the probability weighting function makes low probabilities (close to both 0 and 1) to be over-weighted. The combination of both functions implies a *fourfold pattern* of risk attitudes that is confirmed by experimental evidence: risk aversion for gains and risk seeking for losses of moderate to high probability; risk seeking for gains and risk aversion for losses of low probability.

Prospect theory as initially defined by Kahneman and Tversky (1979), may lead to a *violation of in-betweenness* –a counterintuitive effect where the certainty equivalent of a lottery is not in between the smallest and the largest possible payoff of the lottery. To avoid this, Tversky and Kahneman (1992) suggested CPT –which applies the probability weighting to the cumulative distribution function. Yet, an easier approach –here we follow Hens and Bachmann (2008)- is to simply normalize the decision weights *w*(*p*) so that they add up to 1 and can be interpreted again as a probability distribution. For two-outcome prospects, normalized weights *w**(*p*) are calculated as

$$w^{*}(p) = \frac{w(p)}{w(p) + w(1-p)} \tag{5}$$

where *w**(*p*) means normalized weights according to normalized prospect theory, NPT. This approach, a deterministic NPT, is the one we will follow here.

For elicitation purposes, we are going to use a parametric specification. They are generally less susceptible to response error and more efficient than non-parametric methods (Abdellaoui et al., 2008), in the sense that the latter require more questions to be implemented. This is important for a test that requires to be simple, with a short number of questions, and that seeks to minimize the possible effects of response errors and misunderstanding by the respondents.[3]

Based on extensive literature review, we choose two classic specifications.[4] First, the (piecewise) power function by Tversky and Kahneman (1992),

$$v(x) = \begin{cases} x^{\alpha+} & for\, x \geq 0 \\ -\beta(-x)^{\alpha-} & for\, x < 0 \end{cases} \tag{6}$$

[3] Parametric methods also have their flaws (Abdellaoui et al., 2007; Booij et al., 2010). They depend on the suitability of the selected functional forms –we do not know whether measures are driven by the data or by the imposed parametric model. They also suffer from a *contamination effect*: a misspecification of the utility function will also bias the estimated probability weights and vice versa (Abdellaoui, 2000).

[4] Notwithstanding, the same tests may be used for alternative parametric specifications. We have chosen these two for simplicity, because they are classics in the literature, and following Stott (2006)'s combinatorial approach. Stott has the merit of trying to disentangle the contamination effect of parametric measurements by testing eight value functions in combination with eight weighting functions and four choice functions. His combinatorial approach support the power function form when combined with Prelec-I weighting function and Logit stochastic choice function.

where x accounts for gains (if $x \geq 0$) or losses (if $x \leq 0$), α^+ measures sensitivity to gains, α^- does the same to losses, and β measures loss aversion, is the most widely used parametric family to represent the value function because of its simplicity and its good fit to experimental data (Wakker, 2008). Second, the classic Prelec-I weighting function (Prelec, 1998) given by

$$w(p) = \exp(-(-\log(p))^{\gamma}) \tag{7}$$

where $\gamma>0$, to estimate the probability weighting function, with decision weights $w(p)$ being subsequently normalized to $w^*(p)$ following NPT.

To sum up, we have five parameters (α^+, γ^+, α^-, γ^- and β) that we must estimate. However, we must deal with the problem with loss aversion. A major challenge in the literature on prospect theory is that neither a generally accepted definition of loss aversion, nor an agreed-on way to measure it is available. In regards to the first issue, loss aversion as implicitly defined by Tversky and Kahneman (1992) depends on the unit of payment (Wakker, 2010). Only when the curvature parameters in the power function are the same loss aversion can be a dimensionless quantity. Alternative interpretations of loss aversion were provided (see Booij et al., 2010, for a discussion). However, none of these definitions provide a straight index of loss aversion, but formulate it as a property of the utility function over a whole range.

The second drawback is how to measure loss aversion. A correct measurement requires utility for gains and for losses to be determined simultaneously. Some authors have provided solutions for elicitation of loss aversion (see for instance Abdellaoui et al. 2008, Booij et al. 2010), but the debate is still open. We opt for a solution that is inspired by Booij et al. (2010) –by picking up *“all the questions around the zero outcome”* (p.130)- and by empirical finding that utility is close to linear for moderate amounts of money (Rabin, 2000). What we do is to ask participants for a few prospects with small amounts of money and assume $\alpha^+ = \alpha^- = 1$ to estimate β (as a mean or median). However we are aware this only serves as an imperfect solution to a more complex problem, as an index that is constructed by taking the mean or median values of the relevant values of x is not an arbitrary choice (Booij et al., 2010).

Test design

For parameter estimation, various elicitation procedures have been proposed in the literature. Our method merges some characteristics of Tversky and Kahneman (1992)’s approach to elicit certainty equivalents and Abdellaoui et al. (2008)'s proposal to make an efficient test with a minimum number of questions. In particular, the methodology we use is based on the elicitation of certainty equivalents of prospects with just two outcomes. Following Abdellaoui et al. (2008), the elicitation method consists of three stages, with fifteen questions in total: six questions involving only positive prospects (i.e., a chance to win some positive quantity or zero) to calibrate α^+ and γ^+, six questions for negative prospects to calibrate α^- and γ^-, and three questions regarding the acceptability of mixed prospects, in order to estimate β. Then calibration requires to calibrate jointly α^+ and γ^+ for gains with responses to the first set, and α^- and γ^- for losses with those in the second set, using a nonlinear regression procedure separately for each subject. Finally we estimate β, the loss aversion parameter, using information from the mixed prospects.

Several aspects were considered in all three stages. First, utility measurements are typically of interest only for significant amounts of money (Abdellaoui et al., 2008) while utility is close to linear for moderate amounts (Rabin, 2000). Hence, prospects devised to calibrate α^+*,* γ^+, α^-, and γ^- used significant, albeit hypothetical, amounts of money of 500, 1,000 and 2,000 eur –with all outcomes in euros and multiples of 500 eur to facilitate the task for the subjects (Abdellaoui et al., 2008). Second, only the three questions devised to estimate β used small amounts of money for reasons already described. Consequently, with the aim of preventing the possibility that asking the larger amounts in first order might affect the perception of the smaller amounts in the β elicitation, those three questions were asked in first order. Finally, prior to solving any trial, respondents were asked to answer a practice question to familiarize them with the experimental setting. Instructions emphasized there were no right or wrong answers (Booij et al., 2010), but that completing the questionnaire with diligence, always providing objective and honest answers, was a prerequisite to participate in the strategy game (Peón et al., 2014) they were about to perform in the same session, where they would compete for a prize.

The first three questions, regarding the acceptability of a set of mixed prospects, were then provided to participants in sequential order. Specifically, respondents were asked a classic question (Hens and Bachmann 2008, p.120): *"someone offers you a bet on the toss of a coin. If you lose, you lose X eur. What is the minimal gain that would make this gamble acceptable?"*, where *X* took the values 1 eur, 10 eur and 100 eur in the first, second and third iterations, respectively. Posed this way, all questions to calibrate loss aversion set probabilities of success and failure equal to 50%, p = 0.5. Since $w^*(0,5)$ = 0,5 under NPT, the answer provided by the respondent makes the utility of a gain (V^+) equivalent to the disutility of a loss (V^-). Consequently, for the piecewise power utility function we have

$$\beta = \frac{G^{\alpha+}}{(-L)^{\alpha-}} \tag{8}$$

where G means gains, L losses, and loss aversion would be equal to the ratio $G/|L|$ when $\alpha = \alpha^{+} = \alpha^{-}$ or, in particular, if as we assumed $\alpha^{+} = \alpha^{-} = 1$ for small amounts of money.

In the second stage a set of six questions involving only positive prospects was provided, again in sequential order. Figure 1 shows one of the six iterations participants had to answer. Participants had also time to practice a sample question.

[Insert Figure 1 here]

In every iteration participants had to choose between a positive prospect (left) and a series of positive, sure outcomes (right). Information was provided both in numerical and graphical form. Every time the subject answered whether she preferred the prospect or the sure gain, a new outcome was provided. This process was repeated until the computer informed the respondent that the question was completed and she could continue to another prospect. The probabilities of success in all 6 prospects were different (having 2 questions with probability 50% and one question with probabilities of success 99%, 95%, 5% and 1% each), which was emphasized to participants to avoid wrong answers.

The series of sure outcomes per prospect were removed from two sets, following Tversky and Kahneman (1992) in spirit: the first set logarithmically spaced between the extreme outcomes of the prospect, and the second one linearly spaced between the lowest amount accepted and the highest amount rejected in the first set. All sure outcomes were rounded to a multiple of 5 to facilitate the task. Following Abdellaoui et al. (2008), to control for response errors we repeated the last sure outcome of the first series at the end of each trial, allowing to check for the reliability of the responses. The certainty equivalent of a prospect was then estimated by the midpoint between the lowest accepted value and the highest rejected value in the second set of choices. Tversky and Kahneman (1992) emphasize this procedure allows for the cash equivalent to be derived from observed choices, rather than assessed by the subject.

Finally, the third stage included a set of six questions involving only negative prospects, designed to calibrate α^{-} and γ^{-} parameters. We proceeded similarly. Participants had time to practice a sample question. We emphasized every now and then that prospects and sure outcomes were now in terms of losses. We also emphasized that probabilities were in terms of probabilities of losing, which might be different for each prospect (similar probabilities of failure were provided, namely 1%, 5%, 50%, 50%, 95% and 99%). Certainty equivalents were now estimated as the midpoint between the lowest (in absolute terms) accepted value and the highest (in absolute terms) rejected value in the second set of choices.

3. EXPERIMENTAL RESULTS. A SIMPLICITY – EFFICIENCY ANALYSIS

We organized a series of five experimental sessions that took place in the Faculty of Business and Economics (*Universidade da Coruña*, UDC) during October, 2013. A sample of students of different levels and degrees was selected. To make the call, which was open to the target groups, we got in direct contact with students from UDC during their classes to explain what the experiment would consist of, date and time of the sessions, that they would be invited to a coffee during the performance of the tests, and that one of the tests they would complete consists of a game (Peón et al., 2014) where one of the participants per session would win a prize of 60 euros.[5] In total 126 volunteers, all of them under- and post-graduate UDC students, participated in the experiment. All sessions took place in a computer room; participants in the same session completed all tests at the same time, each respondent in a separate computer.

Before completing the overconfidence tests and the risk profiler for prospect theory, participants signed a consent form and completed a questionnaire on demographic information. This required respondents to declare their (a) gender, (b) age, academic background –about (c) level and (d) degree- and (e) professional experience. Table 1 summarizes these priors and values they may take.

[Insert Table 1 here]

Univariate analysis highlights some pros and cons of our sample. On the negative side, all participants are college students. As a consequence, the sample is limited in terms of age (98.4% of participants were between 17 and 28 years old). Besides, age, academic year (level) and professional experience are correlated. Furthermore, level happens to be not a good proxy for education. For hypothesis testing in the literature, education is intended to

[5] A classic problem of framed field experiments is in regards of their external validity. Most literature on experimental economics considers that the incorporation of incentives improves their validity (see Peón et al., 2014). We incorporated the incentive of a 60 euro prize in the strategy game that participants were about to play in the same session, while they were informed that their right to claim the prize was conditioned to their diligent behavior in the behavioral tests, being objective and honest at all times. They were informed as well that the check questions in the PT test were to be used to identify those participants that were inconsistent in their responses. No winners were eventually penalized.

measure levels such as ‘no education’, ‘primary education’, ‘secondary education’, and so on. In our sample, however, level measures only college years and is highly correlated with age. These problems will represent a drawback for hypothesis testing in section 4. On the positive side, the sample is balanced in terms of gender, as well as in terms of age and academic year within the bounds of our sample. Besides, we considered a subsample of 21 students that have no degree in economic or financial studies to serve as contrast.

In what estimations for the behavioral variables is concerned, Table 2 summarizes the basic univariate statistics. Overprecision measures M_{med} and M_{avg} have 125 observations due to missing responses by one participant at that test.

[Insert Table 2 here]

This section aims to assess the reliability of the parameters that were estimated. For such purpose, we conduct a simplicity – efficiency tradeoff analysis to compare the results in this experiment with regular results in both the theoretical and empirical literature. We conduct this analysis separately for each section.

3.1 Reliability of tests on Overconfidence

We analyze separately the goodness of tests devised to estimate **E** and **P** on one hand, and **M** on the other, since they use different tests.

Trivial tests (indicators **E** *and* **P***)*

Participants completed the four trivia in about 15 minutes, instructions included. There were no relevant incidents in any of the five sessions: respondents declared a perfect understanding of instructions, all responses were coherent and there were no missing values of any kind. Finally, the results obtained support tests were designed satisfactorily for the following reasons.

First, participants on average exhibited overestimation (clearly) and underplacement. The average respondent overestimated her performance by 2.9 right answers (out of 40 questions in total). This bias was persistent in both easy and hard tests. On the other hand, the average respondent considered herself below average by -2.7 correct answers, with the bias being mostly attributable to an underplacement in hard tasks.

These findings are consistent with most literature supporting a general bias towards overestimation of one’s abilities (Lichtenstein et al., 1982; De Bondt and Thaler, 1995;

Daniel et al., 2001) except on easy tasks or in situations where success is likely or individuals are particularly skilled (Moore and Healy, 2008), and a general bias towards underplacing one's performances relative to others on difficult tasks (Moore and Small, 2007) or being generally pessimistic about winning in difficult competitions (Windschitl et al., 2003). Table 3 summarizes average responses (out of 10 questions per trivia).

[Insert Table 3 here]

Second, there is a strong correlation between **E** and **P** (see section 4 for more info). That is, though the biases along the sample are towards overestimation and underplacement, participants that exhibited the highest overestimation tend to consider themselves above average (or, at least, featured a lower underplacement) and vice versa. This finding would support the interpretation of overestimation and overplacement as "interchangeable manifestations of self-enhancement" (Kwan et al., 2004; Moore and Healy, 2008).

Finally, the trivia tests were devised to control for the hard – easy effect. However, that design did not work well enough as results suggest we failed to propose a couple of easy tests that participants find them as easy as we expected. As we may see in Table 3 above, trivia tests T2 and T3 had average (median) correct answers of 2.29 (2.0) and 2.75 (3.0). Correct answers attributable only to good luck would represent a coefficient of 2.0,[6] so it shows participants found these tests hard indeed. Trivia tests T1 and T4, instead, were expected to yield correct answers of 7.0 to 8.0 on average,[7] but respondents only hit the right answer 5.4 (5.0) and 5.58 (6.0) out of 10 questions on average (median). This would represent a couple of tests of a medium –rather than an easy- difficulty for respondents.

In any case, results are good for hard tests and coherent with literature for easy (medium) tests, since overplacement reduces from -2.4 in hard tests to about zero in easy ones, while overestimation does not increase (supporting the finding that a general bias towards overestimation is appreciated). Figure 2 helps to appreciate this effect more clearly.

[Insert Figure 2 here]

Most observations for the hard tests (the graph on the RHS in Figure 2) meet the mentioned tendency towards overestimation and underplacement. For tests with a

[6] Each test consisted of ten questions with five possible answers each. Hence, participants had a probability of 20% to hit the right answer by chance, making it 2.0 right answers out of 10.

[7] Those were the results obtained in a pre-test with similar questions performed by several volunteers. We attribute the eventual differences between the experiment and the pre-test to differences in age and experience between both samples (for instance, volunteers in the pre-test included teachers as well as students, and elder people might had better clues for a right answer in questions about events that happened decades ago). Otherwise, readers may also attribute our failure to researchers' overconfidence.

medium difficulty (the graph on the LHS of Figure 2) the general drift upwards is noticeable (meaning lower levels of underplacement for easy tests are general along the sample), while overestimation is similar on average but with less observations towards higher levels. Furthermore, it is also clear that the correlation between overestimation and overprecision mentioned above exists in both instances.

*Test on confidence intervals (indicator **M**)*

Participants completed the six questions on confidence intervals to infer their individual degree of overprecision (estimator ***M***) in about 6 to 8 minutes, instructions included. Though results show a vast tendency towards overprecision that is supported by most empirical findings in the literature (e.g., Jemaiel et al., 2013), we are concerned about the reliability of the estimations obtained at the individual level. We will later explain why; for now let us analyze the main results obtained.

First, judges were significantly overconfident. The aggregate results show a strong tendency to overprecision: the 80% confidence intervals contained the correct answer only 38.3% of the time. This is much higher than the 14% overconfidence observed by Soll and Klayman (2004) for three-point estimates and about the same level than for a range estimate. Overconfidence varied across domains as it was expected: the lowest degree of overprecision corresponds to the domain where participants could draw on personal experience (time to walk from one place to another). However, they were still overconfident: 80% intervals hit the right answer 62.0% of the time.

When the M ratios are estimated to account for the effects of variability, overprecision becomes even more prevalent: almost 75% of respondents exhibit overprecision ($M < 1$) in the domain with the lowest level ('time to walk') and 97.6% in the highest ('how many deaths'). When these results are added up to calculate a single ratio M per judge, 93.6% (mean) and 97.6% (median) of the respondents exhibit overprecision. Finally, we use Soll and Klayman's alternative refinement to estimate M to see[8] overprecision is mainly attributable to narrow size intervals. Table 4 summarizes all these results.

[Insert Table 4 here]

[8] The original refinement is the one we already explained: doing the estimates of MEAD and MAD based on the beta function that better fits the three point estimations provided by the respondent. Alternatively, Soll and Klayman (2004) suggest we could measure MAD assuming the median is in the middle of the distribution (i.e., using only the two endpoints and assuming a symmetric distribution). The authors denoted M_3 the first measure and M_2 the second one.

As we may see in Table 4, when ratio **M** is estimated assuming the median is in the middle of the distribution rather than using the participant's response (denoted M_2 in the authors' notation) overprecision slightly increases. This means most participants with an asymmetric SPDF tended to provide median estimates that reduced the errors (at least to some extent). This result is coherent with Soll and Klayman's empirical finding that three-point estimates reduce overconfidence.

Although results on aggregate seem to be consistent with empirical literature, we are concerned about the reliability of the estimations obtained at the individual level. In particular, we are concerned for several reasons. First, there is evidence that many participants did not fully understand the instructions. We had several incidents: a respondent with missing responses; some observations where minimum and maximum boundaries were swapped; others where answers were provided in some particular order (e.g., low – medium – high) when they were required as median – lower – higher; and median estimations that were identical to any of the boundaries.[9]

To avoid these incidents in future research, we suggest to enhance Soll and Klayman (2004)'s approach by setting the order of estimates in terms of lower bound – median – upper bound. According to the authors, "if order of estimates has effects, they are complex ones" (p. 311), which supports our suggestion that a specific order will not bias the results but helps respondents to better understand the task. A picture would also be very helpful, such that they are required to fill three boxes in the specific order.

The second reason why we are concerned about reliability of data is because individual estimations of **M** are highly variable depending on the refinement method and whether indicators are estimated as the median or the average of the ratios across domains. In particular, we compared three alternative refinement methods (the two already described and a third one where both MEAD and MAD computations assume a normal distribution), and for each of them we computed the individual indicator **M** as either the mean or median of the ratios across domains. We get the results summarized in Table 5.

[Insert Table 5 here]

First, the last refinement method that assumes normality yields the most extreme results. We will later see this effect is not a problem of this particular method but an evidence of a

[9] Fortunately, we could contact participants by e-mail days after the tests were performed to ask them to confirm their answers. This way, errors of the kind swapped boundaries or responses in a particular order could be amended. Others instead, like missing values or median estimations identical to any boundary, were not modified as it would represent an alteration of the experiment results.

weakness of the test itself. Second, indicators that are computed as average ratios are higher. Third, if we compare how many individuals have an estimator that varies substantially[10] whether we use medians or averages, about half of the individuals have an indicator that is highly sensible to the estimation method. This effect is particularly pervasive when the *M* ratios yield qualitative results that are conflicting: i.e., when we have the same individual could exhibit overprecision (M<1) or underprecision (M>1) depending on the method we consider. This happens to 4% of participants in the standard refinement and up to 9.6% in the worst case. Finally, if we do this comparison across refinement methods[11] (instead of median vs. average) we obtain similar results.

Why this happened? Basically, because in our search for a simplicity – efficiency equilibrium we heeled heavily over simplicity: we designed the tests with only two questions per domain and this revealed to be not enough. If a judge happens to provide an answer to a question that is very close to the true answer, AD will be near to zero. When only having two questions per domain this makes MAD → 0 and **M** → ∞, which would distort our mean estimation **M** across domains –since we only have three. Besides, given the nature of the reliability problem, average estimations tend to be less reliable than median estimations.

Though this effect is more palpable in the case of the refinement method that assumes normality, this only happened by chance. In particular, there were a few respondents (basically only four) for which the middle point of their inferred symmetric SPDF for a particular question happened to be very close to the true answer. Would this happen instead with the median answer provided by the judges, the effect would be more palpable for the original *M* indicator we are using as option-by-default.

3.2 Reliability of tests on Prospect Theory

This section analyzes the reliability of our method to measure the value and weighting functions of a respondent. Participants in the experiment completed the fifteen questions in about 20 minutes, instructions included, and there were no relevant incidents in any of the five sessions.

[10] We consider a 'substantial variation' of 0.10 in absolute terms between median and average estimations. Since median estimations of *M* in the different methods are about 0.40, a variation of 0.10 would represent an estimation that varies about 25% depending on the method we use –which we consider a variation that is substantial enough. Given this variation is basically equivalent to the median variations observed along the sample for the three refinement methods (0.09 – 0.10 according to Table 5), it is not a surprise that we had in all cases about half the individuals affected by a sensible measure.

[11] Given we have three different refinement approaches, we have done this comparison across methods by analyzing the minimum and maximum estimations we get for each individual using any of the three methods.

Results evidence our tests resulted largely satisfactory to replicate the main findings of prospect theory. We support the validity of our method based on several analyses, both at the individual and aggregate level: (i) properties of the value and weighting functions; (ii) the fourfold pattern of risk attitudes; (iii) iteration and fitting errors; (iv) anomalies detected at the individual level. We explain these analyses in detail in what follows.

Value and weighting functions

Results at the aggregate level are described with four measures: the average and median of parameters estimated at the individual level, and the parameters estimated for the average and median participant. Table 6 provides the results at the aggregate level. We also compare our results in Table 6 against some classical results in the literature (where the power and Prelec-I specifications were used).[12]

[Insert Table 6 here]

Most empirical estimations of utility curvature support the assumption of concavity for gains (α^+ from 0.7 to 0.9 in most studies) and convexity for losses (α^- from 0.7 to 1.05), with more recent studies providing estimations closer to linearity in both instances (Booij et al., 2009). Our results reiterate these findings for gains, while risk seeking in the negative domain seems to be more acute (this assertion will be later qualified). The percentage of individuals with alpha measures below one are 59.5% (α^+) and 93.7% (α^-).

We observe a significant degree of probability weighting in both domains –with distortion being higher in the negative side- and quantitative estimations (about $\gamma^+ = 0.6$ and $\gamma^- = 0.5$) are in consonance with literature. By using Prelec-I function we are imposing the classic inverse S-shaped weighting function observed in most studies (that is, the non-linear regressions set the restrictions $\gamma \leq 1$). Notwithstanding, there seems to be no debate here since aggregate indicators are significantly below 1 and most individual observations (78% for gains, 91% for losses) fitted better for gamma values below 1.[13]

Parameters α^- and γ^- suggest a strong risk seeking behavior in the negative domain by most participants. There may be two interpretations that are not mutually exclusive. Results might suggest most participants were unable to fully interpret hypothetical losses as real. In particular, several participants were strongly biased in terms of probability

[12] Results provided for comparison include Tversky and Kahneman (1992), Abdellaoui et al. (2007), Abdellaoui et al. (2008), Wu and Gonzalez (1996), Stott (2006), Bleichrodt and Pinto (2000), Donkers et al. (2001) and Booij et al. (2009). More information about other authors, as well as results for other parametric specifications, are available in extensive summaries provided by Stott (2006) and Booij et al. (2009).

[13] We computed all respondents with γ^+, $\gamma^- < 0.95$.

weighting (the minimum observation is $\gamma^{-} = 0.05$ and one third of the sample is below the lower bound in the literature, 0.35) and most of them exhibited a utility curvature α^{-} below 0.50. Second, some individuals' profile might be better described with a weighting function that accounts for elevation as well as curvature, like Prelec-II (see 'anomalies at the individual level' for more info). Besides, a contamination effect might also affect α^{-} estimations –which are below regular results in the literature, as we noted.

Finally, our beta estimations are in consonance with classic results in the literature (a loss aversion higher than 2) compared to more moderate estimations reported by Booij et al. (2009). The percentage of individuals with beta measures above two are 73.0% for β_{med} (65.7% for β_{avg}) and only 14.3% have $\beta_{med} \leq 1$ (7.9% using β_{avg}).

The fourfold pattern of risk attitudes

Tversky and Kahneman (1992) analyze the fourfold pattern of risk attitudes by plotting, for each positive prospect of the form (x , p ; 0 , 1-p), the ratio of the certainty equivalent c of the prospect to the nonzero outcome x, c/x, as a function of p. We do the same in the negative domain, so we get two different graphs of c/x over p. Figure 3 provides these plots for the certainty equivalents provided by the average (idealized) participant.

[Insert Figure 3 here]

Should we estimate two smooth curves, one per domain, they would be interpreted as weighting functions assuming a linear value function. The fourfold pattern of risk attitudes in prospect theory predicts we tend to be risk seeking for gains of low probability (1% and 5% in our test) and losses of medium and high probability, while we tend to be risk averse for gains of medium and high probability and losses of low probability. The pattern is clearly observable for the average respondent, with the nuance of an about risk neutrality for gains of medium probability. Results for the median respondent are quite similar.

When we extend this analysis to the individual level we get the results summarized in Table 7.[14] The risk attitudes predicted by prospect theory in the positive domain are generally satisfied, with about 2/3 of the elicitations being risk seeking for low probabilities and risk averse otherwise. In the negative domain the bias towards risk seeking is more evident, making results for low probabilities mixed.

[Insert Table 7 here]

[14] For risk-neutrality in Table 7 we report the percentage of elicitations that revealed a certainty equivalent that was the closest possible to the expected value of the game.

Iteration and fitting errors

We determine the validity of participants' responses based on two kinds of errors. The first type, iteration errors, refers to the reliability of the iterative questions we asked to control for response errors. The second type, fitting errors, refers to those obtained in the non-linear regressions implemented for parameter estimation assuming the pre-specified parametric forms.

Abdellaoui et al. (2008) argue that one of the main strengths of their model is that by allowing for response error during the elicitation process, the number of questions required to measure the value function is minimized. In particular, they repeated two types of iterations[15] to obtain 96% reliability for the first replication and 66% for the second one, and they claim them to be satisfactory. Using a similar approach, we repeated one iteration per question (with a somehow similar interpretation to Abdellaoui et al.'s second replication) for all twelve questions in the positive and negative domains. The results were highly satisfactory: on aggregate, only 5.6% of responses were contradictory (94.4% reliability). Furthermore, 65.4% of participants made not a single response error, 81.7% made one error at most, and only 2 out of 126 participants made more than three.

These results confirm that the experiment design (graphics, instructions and practice questions) was helpful for participants to correctly understand the task. Whether some risk profiles are not common (as it happens with α^- and γ^- estimations noted above), it may hence be attributed to the difficulties for some participants to imagine hypothetical losses as real, but not to a misinterpretation of data.

In what fitting errors is referred, the high quality of the R^2 coefficients obtained to estimate the PT parameters for most individuals are both an additional confirmation that participants understood the task, as well as an indicative that the parametric functions we used were satisfactory. For those respondents whose coefficients were low, this in most cases might only indicate that with other value and/or weighting functions the fitting quality would improve. Nonetheless, in section 'anomalies at the individual level' below we analyze some results that are difficult to rationalize and that might reveal some mistakes or confusion by the respondent. Table 8 summarizes the R^2 obtained.

[Insert Table 8 here]

[15] Abdellaoui et al. (2008) repeated "the first iteration after the final iteration" for all questions, and "the third iteration" of 2 questions for gains and 2 for losses, chosen randomly.

Results are slightly better in the positive domain, with about 80% and 65% of individual regressions being satisfactory and only three observations (2.4%) in the positive domain and one (0.8%) in the negative domain being really weak.

Anomalies at the individual level

The coefficients of determination R^2 are helpful to identify some results at the individual level that are difficult to put in consonance with the basic predictions of prospect theory. We highlight eight cases[16] whose risk attitudes (plotting of *c/x* over *p*) are described in Figure 4. It seems difficult not to agree some answers reveal a response error. To illustrate, p = 0.99 in the positive domain of case 4 or the same probability in the negative domain of case 6. Other examples reveal profiles that are hard to rationalize. Take for instance case 7 in the positive domain (the lowest coefficient of determination, $R^2 = 0.05$), where the respondent required 355 euros for not accepting a prospect to win 1,000 euros with 5% probability, but a lower amount (342.5 euros) for not accepting 2,000 euros with p = 95%. Similar situations appear when comparing responses for p = 50% with high and low probabilities (e.g., case 1 in the negative domain or 8 in the positive one).

[Insert Figure 4 here]

However, some other cases might reveal a risk profile that is too aggressive or unusual, but not necessarily a response error. Take for instance case 3 in the negative domain, which features a high risk seeking profile, or cases 2 and 6 in the positive domain, which might reveal that the inverse-S shaped weighting function is not suitable for them. Consequently, we conclude we cannot detect anomalies based solely on R^2.

4. HYPOTHESIS TESTING

This section aims to test the effect of priors over behavioral variables, and the relationship among variables. Regarding the first effect, we test the following hypothesis, based on extensive literature review. First, women are (i) less overconfident than men (Lundeberg et al., 1994; Kuyper and Dijkstra, 2009), (ii) exhibit a larger degree of loss aversion (Schmidt and Traub, 2002; Booij et al., 2010), and (iii) are more risk averse in terms of utility curvature and weighting function (Booij et al., 2010). Second, age (iv) reduces

[16] These individuals show the lowest fitting accuracy on any of the two domains or both. As an additional piece of evidence, all but one of these individuals made at least one iteration error, for an average of 1.75 errors per respondent. A performance that is statistically higher ($p < 0.01$) than the 0.61 mean error of all the other participants, suggesting these profiles correspond to judges that had more problems to understand the task.

overconfidence (Sandroni and Squintani, 2009; Zell and Alicke, 2011). Third, education or, alternatively, working experience (v) induces a more linear probability weighting (Booij et al., 2010 find evidence against this), (vi) reduces loss aversion (Donkers et al., 2001) and (vii) moderates both over- and underconfidence. Fourth, we hypothesize that skills in finance (viii) reduce overconfidence, (ix) increase loss aversion, (x) increase risk aversion, and (xi) induce a more objective (linear) probability weighting.

Regarding the relationship among variables, we trace relationships of three kinds: among different overconfidence measures, among prospect theory parameters, and between overconfidence and prospect theory. This is also a main contribution of our research, since we are not aware of previous works where these two relevant areas of behavioral finance were analyzed simultaneously with a same sample. Prior to solve the hypothesis testing, normality tests and box plots were used to remove four observations from two variables, one extreme value for age and three for loss aversion (β_{avg}). We test the hypotheses with two alternative methods, a correlation analysis and a regression analysis. The most relevant results we obtain are in order.

Regarding priors and variables, a significant correlation appears between level and loss aversion ($p < 0.05$), but with a positive sign, rejecting the null hypothesis in test (vi). Despite these results, we declared level in our sample to be a bad proxy for education, so we would take the interpretation that education increases loss aversion only carefully.[17] We also find evidence ($p < 0.05$) that experience reduces objectivity in terms of estimation of self-performance –contrary to hypothesis (vii).

In regards to statistical correlation among behavioral biases, more relevant results appear. There is evidence that overestimation and overplacement are correlated ($p < 0.01$), but we do not support Moore and Healy (2008)'s assertion that overprecision reduces them both. Correlation among PT parameters also suggest very interesting results. First, risk seeking comes together in both domains: α^+ and α^- are negatively correlated ($p < 0.05$). Second, objective weighting of probabilities also come together in both domains: γ^+ and γ^- are positively correlated ($p < 0.01$). Finally, there is strong evidence that loss aversion and risk aversion in the negative domain come together as well.

[17] Several other relationships between priors and variables satisfy the null hypotheses to be tested, but with no statistical significance at all. First, age reduces overconfidence: older students exhibit lower levels of overestimation and overplacement (with no statistical significance) as well as of overprecision, with a statistical significance that improves for both measures, but only to about 20%. Second, educated (level) and more experienced individuals (working experience) weight probabilities more linearly, but only in the positive domain (γ^+). Third, working experience reduces (both measures of) loss aversion

Regarding the relationship between overconfidence and PT parameters, we find only positive correlations ($p < 10\%$) between α^- and **E,** and between γ^- and **M**. They are harder to interpret, as they suggest individuals with a more aggressive profile for losses (higher risk seeking and distortion of probabilities) would be correlated with lower levels of overconfidence (in terms of overestimation and overprecision). Further experimental research must determine whether these correlations are spurious.

Hypotheses on gender and skills were tested with an ANOVA test. Regarding gender, women appear to be significantly more overconfident than men in terms of overprecision, contrary to hypothesis (i), more risk seeking[18] both in the positive and negative domain (the latter means women are more averse to a sure loss), contrary to hypothesis (iii), and with a significantly higher distortion of probabilities in the negative domain. Regarding skills in finance, it increases objectivity reducing probability distortion ($p < 0.01$) and reduces risk aversion ($p < 0.1$), both in the positive domain.[19] The first result supports hypothesis (xi) while the second one goes against (x).

The regression analysis yields results that are coherent with correlations. We regress behavioral biases over priors, with gender and skills as dummy variables, and to avoid multicollinearity we perform a stepwise procedure for variable elimination. The models predict women exhibit more overprecision (lower M_{avg}), higher risk seeking in terms of utility curvature (higher α^+ and lower α^-) and higher distortion of probabilities in the negative domain (lower γ^-) than men. Skills in finance explain a more objective weighting of probabilities (higher γ^+) while the more education (level) the higher loss aversion (β_{avg}). The explanatory power of these models is very low in all instances, but significantly different from zero in any case.

5. CONCLUDING REMARKS

We have introduced a set of simple tests to elicit the three measures of overconfidence as well as the complete set of parameters of value and weighting functions in prospect theory. We also provide extensive evidence that the experimental research implemented

[18] Recall we are working under a ceteris paribus condition: the fourfold pattern of risk attitudes requires risk aversion and risk seeking to be discussed in terms of value and weighting functions simultaneously.

[19] Several other relationships satisfy the null hypotheses to be tested, but with no statistical significance at all. Regarding gender these include men are more overconfident in terms of *M* and *P*, while women are more risk seeking (both domains) in terms of utility curvature but more loss averse. Regarding skills in finance, these include reducing overestimation and increasing loss aversion (β_{med}).

to validate our tests confirm they are generally efficient to replicate the standard results in the literature.

In particular, with only four trivia similar to those by Moore and Healy (2008) we obtain satisfactory results in terms of simplicity (it requires only about 8 minutes per indicator) and efficiency to provide individual measures of overestimation and overplacement. A test of fifteen questions in about 20 minutes revealed efficient as well to replicate the main findings of prospect theory, considering the properties of the value and weighting functions, the fourfold pattern of risk attitudes, iteration and fitting errors, and anomalies at the individual level. Our test for overprecision, instead, revealed incomplete to obtain individual estimations that are stable for different refinement methods. In future research, having more questions per domain will be necessary, while it would also be desirable to ask additional questions on personal experience to balance domains.

We are aware of the limitations simplicity induces for elicitation of psychological profiles. However, the main contribution of this paper is to provide a set of tests that are able to obtain efficient results while enhancing the scope for empirical application of prospect theory and overconfidence. The paper also contributes to provide additional evidence about how gender, education and skills in finance affect overconfidence and risk aversion. In particular, the experimental analysis of all measures of overconfidence and prospect theory using the same sample of respondents is something that, to the best of our knowledge, was not done before. This allows us to provide new insight on the relationship between these two relevant areas in the behavioral literature.

Additional enhancements for future research might be introducing questions on abilities and perceptual tasks (Stankov et al., 2012) in the trivia test to moderate the general drift towards overestimation, and setting the computer application in the PT test to refine answers that might be interpreted as a response error by asking an additional questions. Finally, two open questions in the PT test are how to improve loss aversion estimations, since sensibility of the value function to lower amounts of money varies across individuals, and how to foster more realistic answers, particularly in the negative domain as incentives would be an implausible solution as it would require a sample of individuals willing to participate in an experiment where they are offered to lose real money.

REFERENCES

Abdellaoui, M. (2000), Parameter-free elicitation of utility and probability weighting functions, *Management Science* 46(11), 1497-1512.

Abdellaoui, M., H. Bleichrodt and C. Paraschiv (2007), Loss aversion under prospect theory: A parameter-free measurement, *Management Science* 53(10), 1659-1674.

Abdellaoui, M., H. Bleichrodt and O. L'Haridon (2008), A tractable method to measure utility and loss aversion under prospect theory, *Journal of Risk and Uncertainty* 36, 245-266.

Bleichrodt, H. and J.L. Pinto (2000), A parameter-free elicitation of the probability weighting function in medical decision analysis, *Management Science* 46(11), 1485–1496.

Booij, A.S., B.M.S. van Praag and G. van de Kuilen (2010), A parametric analysis of prospect's theory functionals for the general population, *Theory and Decision* 68, 115-148.

Bowman, D., M. Minehart and M. Rabin (1999), Loss aversion in a consumption-savings model, *Journal of Economic Behavior and Organization* 38, 155-178.

Daniel, K.D., D. Hirshleifer and A. Subrahmanyam (2001), Overconfidence, arbitrage and equilibrium asset pricing, *The Journal of Finance* 56, 921-965.

De Bondt, W.F.M. and R.H. Thaler (1995), Financial Decision-Making in Markets and Firms: A Behavioral Perspective, In R. Jarrow et al. (eds) *Handbooks in Operations Research and Management Science*, vol. 9, Amsterdam: Elsevier:385–410.

Donkers, A.C.D., B. Melenberg and A.H.O. van Soest (2001), Estimating risk attitudes using lotteries: a large sample approach*, Journal of Risk and Uncertainty* 22, 165-195.

Gonzalez, R. and G. Wu (1999), On the shape of the probability weighting function, *Cognitive Psychology* 38(1), 129-166.

Hens, T. and K. Bachmann (2008): *Behavioural finance for private banking*, John Wiley & Sons Ltd.

Jemaiel, S., C. Mamoghli and W. Seddiki (2013), An experimental analysis of over-confidence, *American Journal of Industrial and Business Management* 3, 395-417.

Kahneman, D. and A. Tversky (1979), Prospect Theory: an analysis of decision under risk, *Econometrica* 47(2), 263-291.

Köbberling, V. and P. Wakker (2005), An index of loss aversion*, Journal of Economic Theory* 122, 119-131.

Kuyper, H. and P. Dijkstra (2009), Better-than-average effects in secondary education: A 3-year follow-up, *Educational Research and Evaluation* 15(2), 167-184.

Kwan, V.S.Y., O.P. John, D.A. Kenny, M.H. Bond and R.W. Robins (2004), Reconceptualizing individual differences in self-enhancement bias: An interpersonal approach, *Psychological Review* 111(1), 94-110.

Lichtenstein, S., B. Fischhoff and L.D. Phillips (1982), Calibration of probabilities: The state of the art to 1980, In D. Kahneman, P. Slovic and A. Tversky, *Judgement under uncertainty: heuristics and biases*, Cambridge University Press.

Lundeberg, M.A., P.W. Fox and J. Punccohar (1994), Highly confident but wrong: Gender differences and similarities in confidence judgments, *Journal of Educational Psychology* 86, 114-121.

Moore, D.A. and P.J. Healy (2008), The trouble with overconfidence, *Psychological Review* 115(2), 502–517.

Moore, D.A. and D.A. Small (2007), Error and bias in comparative social judgment: On being both better and worse than we think we are, *Journal of Personality and Social Psychology* 92(6), 972-989.

Neilson, W.S. (2002), Comparative risk sensitivity with reference-dependent preferences, *Journal of Risk and Uncertainty* 24, 131-142.

Peón, D., M. Antelo and A. Calvo (2014), Overconfidence and risk seeking in credit markets: An experimental game, *Mimeo.* Available upon request to the authors.

Prelec, D. (1998), The probability weighting function, *Econometrica* 66(3), 497-527.

Rabin, M. (2000), Risk aversion and expected-utility theory: A calibration theorem, *Econometrica* 68(5), 1281-1292.

Rötheli, T.F. (2012), Oligopolistic banks, bounded rationality, and the credit cycle, *Economics Research International*, vol. 2012, Article ID 961316, 4 pages, 2012. doi:10.1155/2012/961316.

Sandroni, A. and F. Squintani (2009), Overconfidence and asymmetric information in insurance markets, *unpublished WP* available at www.imtlucca.it/whats_new/_seminars_docs/000174-paper_Squintani_April6.pdf

Schmidt, U. and S. Traub (2002), An experimental test of loss aversion, *Journal of Risk and Uncertainty* 25, 233-249.

Shefrin, H. (2008), *Ending the management illusion: How to drive business results using the principles of behavioral finance*, Mc-Graw Hill, 1st edition.

Soll, J.B. and J. Klayman (2004), Overconfidence in interval estimates, *Journal of Experimental Psychology: Learning, Memory and Cognition* 30(2), 299-314.

Stankov, L., G. Pallier, V. Danthiir and S. Morony (2012), Perceptual underconfidence: A conceptual illusion?, *European Journal of Psychological Assessment* 28(3), 190-200.

Stott, H.P. (2006), Cumulative prospect theory's functional menagerie, *Journal of Risk and Uncertainty* 32, 101-130.

Tversky, A. and D. Kahneman (1992), Advances in prospect theory: Cumulative representation of uncertainty, *Journal of Risk and Uncertainty* 5(4), 297-323.

Wakker, P.P. (2008), Explaining the characteristics of the power (CRRA) utility family, *Health Economics* 17, 1329-1344.

Wakker, P.P. (2010), *Prospect Theory for Risk and Ambiguity*, Cambridge University Press.

Wakker, P.P. and A. Tversky (1993), An axiomatization of cumulative prospect theory, *Journal of Risk and Uncertainty* 7, 147-176.

Windschitl, P.D., J. Kruger and E. Simms (2003), The influence of egocentrism and focalism on people's optimism in competitions: When what affects us equally affects me more, *Journal of Personality and Social Psychology* 85(3), 389-408.

Wu, G. and R. Gonzalez (1996), Curvature of the probability weighting function, *Management Science* 42(12), 1676-1690.

Zell, E. and M.D. Alicke (2011), Age and the better-than-average, *Journal of Applied Social Psychology* 41(5), 1175-88.

IBM SPSS Statistics version 21 and R Project (Packages rriskDistributions and zipfR) were used for statistical analysis.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Belgorodski, N., M. Greiner, K. Tolksdorf and K. Schueller (2012), rriskDistributions: Fitting distributions to given data or known quantiles. R package version 1.8. http://CRAN.R-project.org/package=rriskDistributions

Evert, S. and M. Baroni (2007), zipfR: Word frequency distributions in R, In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions, 29-32. (R package version 0.6-6 of 2012-04-03)

FIGURES AND TABLES

FIGURE 1 – A sample question with positive prospects

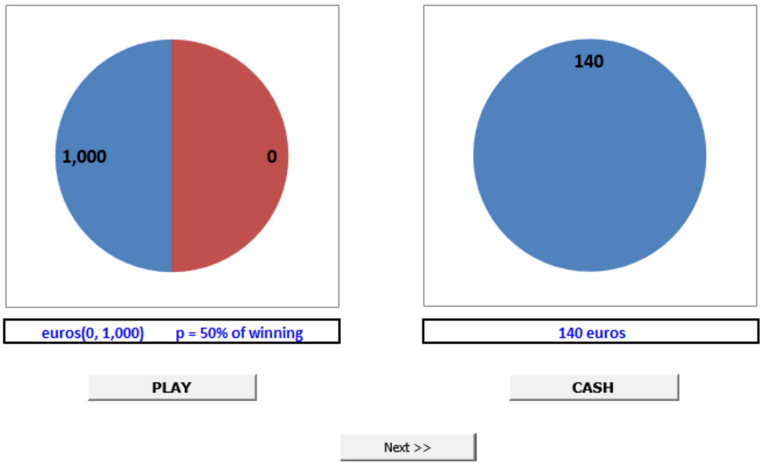

FIGURE 2 – The hard – easy effect

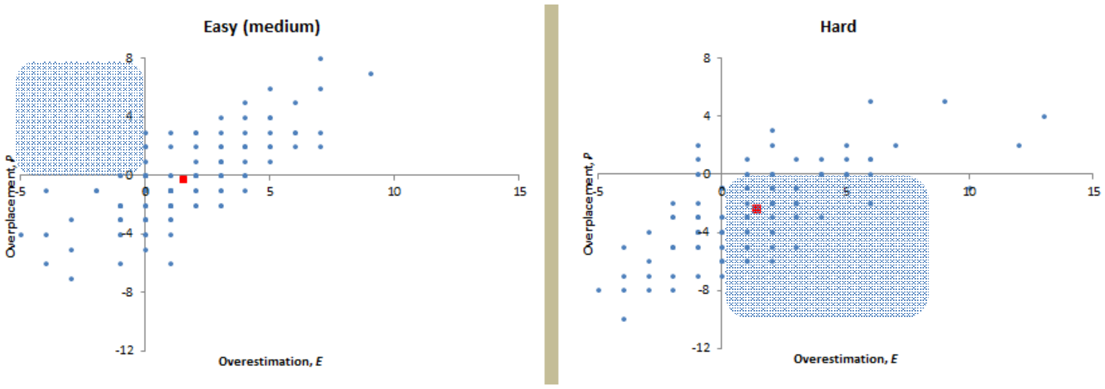

FIGURE 3 – Risk attitudes of the average participant

	Q6	Q5	Q1	Q2	Q4	Q3
p	0.01	0.05	0.5	0.5	0.95	0.99
c/x	0.037	0.121	0.518	0.493	0.769	0.859
	risk seeking	risk seeking	risk seeking	risk averse	risk averse	risk averse

	Q8	Q7	Q9	Q10	Q12	Q11
p	0.01	0.05	0.5	0.5	0.95	0.99
c/x	0.018	0.063	0.267	0.220	0.620	0.691
	risk averse	risk averse	risk seeking	risk seeking	risk seeking	risk seeking

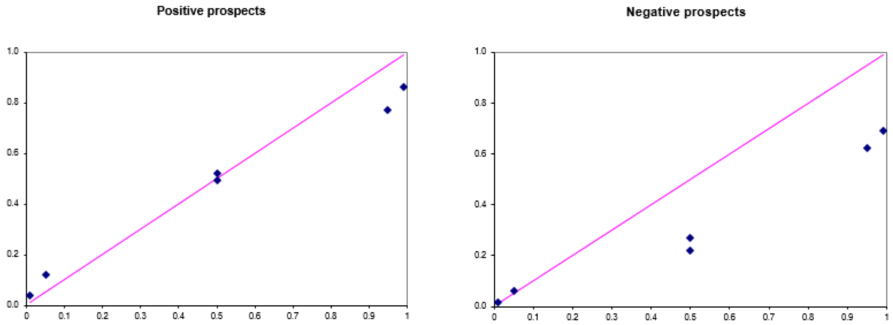

FIGURE 4 – Risk attitudes of eight individual anomalies

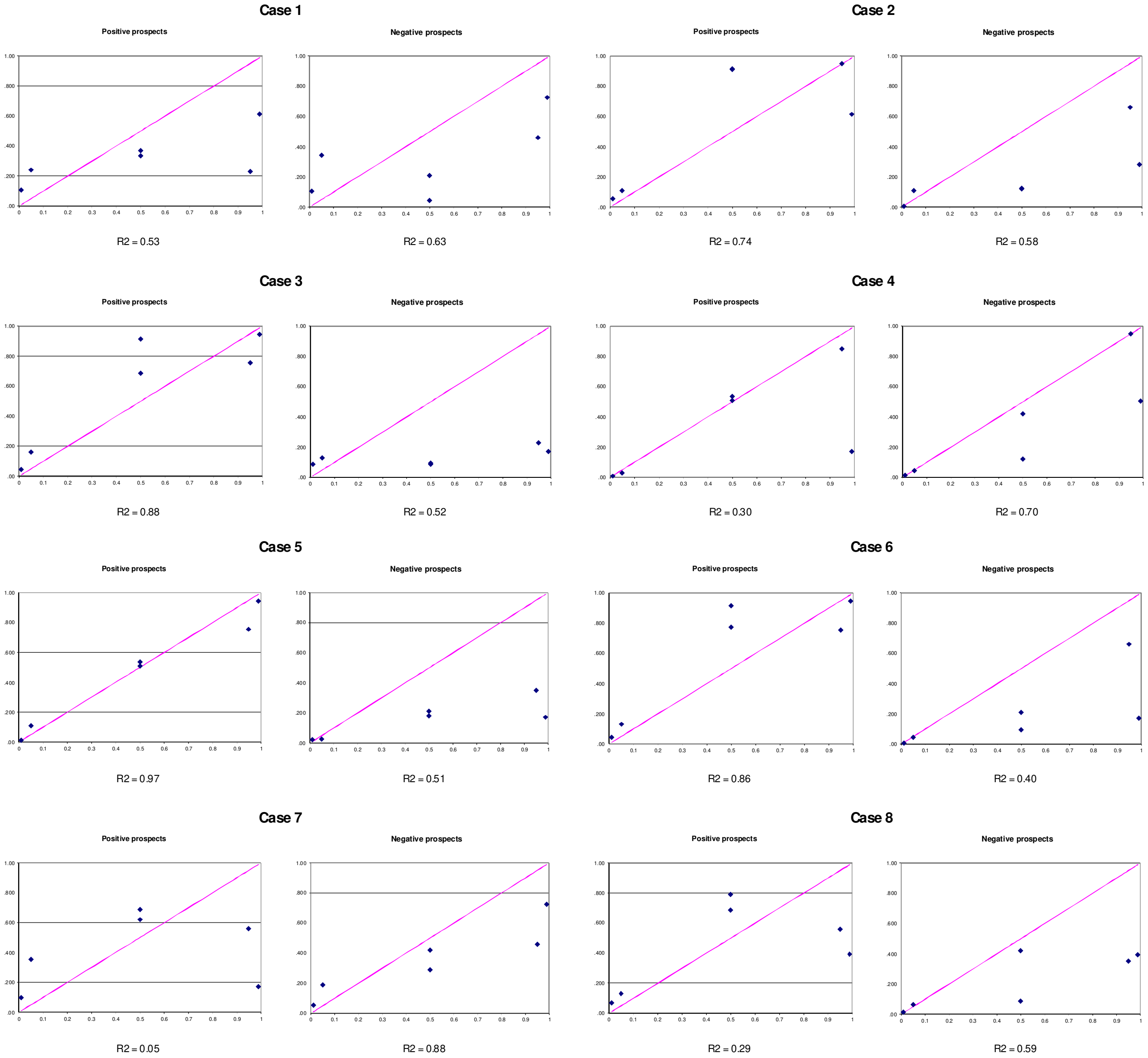

TABLE 1 – Summary of priors

	Variable	Measure	Values
Priors	**Gender**	Nominal	1 = woman; 2 = man
	Age	Scale	# of years
	Level	Scale	1.0 = "1st year"; 2.0 = "2nd year"; ...; 6.0 = "6th year"; 7.0 = "Master of Science, MSc"
	Faculty *	Ordinal	1.0 = "Business and Economics (UDC)"; 2.0 = "Computing"; 3.0 = "Education"; 6.0 = "Law"
	→ ***Skills*** **	*Nominal*	*1.0 = "Others"; 2.0 = "Economics and Business"*
	Experience ***	Ordinal	1.0 = "no experience"; 2.0 = "university trainée"; 3.0 = "occasional employment"; 4.0 = "regular employm."

* Values 4.0 = "Business and Economics (USC)" and 5.0 = "Philology" were initially considered but eventually deleted as we had no observations

** This prior was not directly asked for in the questionnaires but codified using information from 'Faculty'

*** "Ocassional employment" was codified in the questionnaire as working experience with salary lower than 1,000 eur, and "regular employment" otherwise

TABLE 2 – Descriptive statistics of behavioral variables

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Age	126	36.00	17.00	53.00	22.15	3.72	13.825	4.704	.216	37.433	.428
Level	126	6.00	1.00	7.00	4.04	2.22	4.918	.155	.216	-1.400	.428
E	126	28.00	-8.00	20.00	2.93	4.76	22.643	.790	.216	1.529	.428
P	126	27.00	-13.98	13.02	-2.71	4.69	21.959	.302	.216	.785	.428
M_{med}	125	1.50	0.00	1.50	0.34	0.26	.066	1.841	.217	4.902	.430
M_{avg}	125	1.32	0.07	1.38	0.46	0.29	.085	1.310	.217	1.837	.430
alpha +	126	2.43	0.24	2.67	1.02	0.46	.213	1.513	.216	2.482	.428
alpha -	126	2.24	0.05	2.29	0.52	0.31	.098	2.320	.216	9.199	.428
gamma +	126	0.95	0.05	1.00	0.64	0.26	.065	-.163	.216	-.700	.428
gamma -	126	0.95	0.05	1.00	0.53	0.28	.077	.183	.216	-1.147	.428
β_{med}	126	9.40	0.60	10.00	3.01	1.97	3.897	1.599	.216	3.182	.428
β_{avg}	126	26.00	0.67	26.67	3.64	3.57	12.750	3.978	.216	20.157	.428

TABLE 3 – Overestimation and overplacement

	T1	T2	T3	T4
self estimation (average)	6,6	2,7	3,8	5,9
self estimation (median)	7,0	2,5	4,0	6,0
estimation of others (average)	6,4	4,0	4,8	6,4
estimation of others (median)	6,0	4,0	5,0	6,0
right answers (average)	**5,40**	**2,29**	**2,75**	**5,58**
right answers (median)	**5,0**	**2,0**	**3,0**	**5,0**

	ALL	Easy	Hard
Overestimation	**2,9**	1,5	1,4
Overplacement	**-2,7**	-0,3	-2,4

TABLE 4 – Overprecision

Domain	Hit rate*		*M*	*"M_2"*
Invention dates		Invention dates		
Q1	12.0%	median	0.28	0.26
Q2	51.2%	average	0.36	0.37
Average	**31.6%**	M < 1 (%)	**94.4%**	**94.4%**
Number of deaths		Number of deaths		
Q3	17.6%	median	0.10	0.10
Q4	24.8%	average	0.21	0.17
Average	**21.2%**	M < 1 (%)	**97.6%**	**99.2%**
Walk times		Walk times		
Q5	66.4%	median	0.64	0.58
Q6	57.6%	average	0.82	0.81
Average	**62.0%**	M < 1 (%)	**74.4%**	**79,2%**
MEDIAN		M < 1 (%)	**93.6%**	**98.4%**
AVERAGE	**38.3%**	M < 1 (%)	**97.6%**	**96.8%**

* Answers that exactly matched an endpoint were counted as correct

TABLE 5 – Reliability of individual M estimations

	M_{beta}		M_2		M_{normal}	
	M_{med}	M_{avg}	M_{med}	M_{avg}	M_{med}	M_{avg}
range	0.0 - 1.5	0.07 - 1.38	0.0 - 1.59	0.05 - 3.08	0.02 - 4.89	0.08 - 19.68
median	0.31	0.40	0.3	0.38	0.40	0.51
average	0.34	0.46	0.34	0.45	0.51	0.94

		M_{beta}	M_2	M_{normal}
med vs avg	variation*	0.10	0.09	0.09
	threshold**	52.0%	47.2%	45.6%
	change sign***	4.0%	2.4%	9.6%

		median	average
across meth.	variation*	0.09	0.12
	threshold**	46.4%	54.4%
	change sign***	4%	12.8%

* measured as the median of the individual variations

** percentage of individuals for which the difference (in absolute terms) between median and average estimation of M are larger than 0.10

*** percentage of individuals for which ratio *M* ranks the same individual as being both over- and underconfident depending on whether we use median or average estimations

TABLE 6 – PT parameters at the aggregate level

	individual parameters		idealized participant		Main results in the literature*	
	median	average	median	average		
α^+	0.93	1.02	0.96	0.91	- T& K'92: $\alpha^+ = 0.88$ - Abd'08 review : 0.70 to 0.90 - Abd'08 results: $\alpha^+ = 0.86$ - Abd'07: $\alpha^+ = 0.72$	- W&G'96: $\alpha^+ = 0.48$ - Stott'06: $\alpha^+ = 0.19$ - Donk'01: $\alpha^+ = 0.61$ - Booij'09: $\alpha^+ = 0.86$
α^-	0.44	0.52	0.43	0.50	- T& K'92: $\alpha^- = 0.88$ - Abd'08 review : 0.85 to 0.95 - Abd'08 results: $\alpha^- = 1.06$	- Abd'07: $\alpha^- = 0.73$ - Donk'01: $\alpha^- = 0.61$ - Booij'09: $\alpha^- = 0.83$
γ^+	0.63	0.64	0.60	0.52	- T& K'92: $\gamma^+ = 0.61$** - Abd'08: $\gamma^+ = 0.46 - 0.53$ - W&G'96: $\gamma^+ = 0.74$	- Stott'06: $\gamma^+ = 0.94$ - B&P'00: $\gamma^+ = 0.53$ - Donk'01: $\gamma^+ = 0.413$
γ^-	0.50	0.53	0.58	0.40	- T& K'92: $\gamma^- = 0.69$** - Abd'08: $\gamma^- = 0.34 - 0.45$	- Donk'01: $\gamma^- = 0.413$
β_{med}	2.00	3.01	2.00	3.04	- T& K'92: $\beta = 2.25$ - Abd'08 review : 2.24 to 3.01 - Abd'08 results: $\beta = 2.61$	- Abd'07: $\beta = 2.54$ - Booij'09 review : 1.38 to 1.63 - Booij'09 results: $\beta = 1.6$
β_{avg}	2.67	3.64	2.33	3.51		

* Authors mentioned: T& K'92 (Tversky and Kahneman, 1992); Abd'08 (Abdellaoui et al., 2008); Abd'07 (Abdellaoui et al., 2007); W& G'96 (Wu and Gonzalez, 1996); Stott'06 (Stott, 2006); B& P'00 (Bleichrodt and Pinto, 2000); Donk'01 (Donkers et al., 2001); Booij'09 (Booij et al., 2009)

** Results are not comparable as authors imposed a different parametric specifications other than Prelec-I for the weighting function

TABLE 7 – The fourfold pattern at the individual level

	GAINS						LOSSES					
	low		medium - high				low		medium - high			
	p = .01	p = .05	p = .50	p = .50	p = .95	p = .99	p = .01	p = .05	p = .50	p = .50	p = .95	p = .99
risk seeking	63.5%	65.1%	30.2%	21.4%	0.0%	0.0%	47.6%	42.1%	84.1%	88.9%	89.7%	100%
risk neutral	10.3%	16.7%	34.1%	32.5%	15.9%	0.0%	14.3%	19.0%	11.9%	8.7%	10.3%	0.0%
risk averse	26.2%	18.3%	35.7%	46.0%	84.1%	100%	38.1%	38.1%	4.0%	8.7%	0.0%	0.0%

	low	medium - high	low	medium - high
risk seeking	64.3%	12.9%	44.8%	90.7%
risk neutral	13.5%	20.6%	16.7%	7.7%
risk averse	22.2%	66.5%	38.1%	3.2%

TABLE 8 – Coefficients of determination

	positive domain	negative domain
$R^2 \geq 99$	19.8%	19.0%
$R^2 \geq 90$	79.4%	65.1%
$R^2 < 50$	2.4%	0.8%