A short test for overconfidence and prospect theory. An 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 test for overconfidence and prospect theory: An experimental validation*

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

This version: August 2014

Abstract

Two relevant areas in the behavioral economics are prospect theory and overconfidence. Many tests are available to elicit their different manifestations: utility curvature, probability weighting and loss aversion in prospect theory; 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. In this paper we provide two short tests, based on classic works in the literature, to derive a complete profile on prospect theory and overconfidence. Then, we conduct an experimental research to validate the tests, revealing they are broadly efficient to replicate the regular results in the literature. Nonetheless, some enhancements are suggested as well. Finally, 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.

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. Manel Antelo acknowledges financial aid from the Galician autonomous government (Xunta de Galicia) through the project Consolidación e estruturación GPC GI-2060 Análise económica dos mercados e institucións (AEMI).

[a] Grupo BBVA and Departament of Financial Economics and Accountancy, University of A Coruna, Campus de Elviña s/ n, 15071 A Coruña. Email: *david.peon@udc.es*

[b] Corresponding author.

[c] Departament of Financial Economics and Accountancy, University of A Coruna, Campus de 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 on some classic tests in the literature. However, trying to replicate them completely would be unfeasible in terms of time to be performed. Only to illustrate, the classic work by Tversky and Kahneman (1992) reports 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 tests for overconfidence by Moore and Healy (2008) spent "about 90 minutes in the laboratory" to complete 18 rounds of 10-item trivia quizzes. We need shorter tests for our experiment, 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 literature (e.g., Abdellaoui et al., 2008), since they would enhance the scope for application of behavioral theories.

Thus, the objective of this research is twofold. Firstly, the article is devoted to explain how we devised some short tests to obtain a basic profile, in terms of prospect theory and overconfidence, of a given individual, and the literature that supports our choices. Thus, on one hand we follow Moore and Healy's (2008) theory on the three different measures of overconfidence, and design shorter versions of Soll and Klayman's (2004) and Moore and Healy's (2008) tests to elicit those measures at the individual level. On the other, in regards to prospect theory, we follow Rieger and Wang's (2008) normalization of prospect theory (Kahneman and Tversky, 1979) assuming classic parametric functions in the literature, while for test design we merge some features of Kahneman and Tversky's (1992) elicitation method and the approach to make an efficient test with a minimum number of questions by

Abdellaoui et al. (2008). The second objective of this research is to validate the tests devised. To such purpose, they were implemented to a sample of 126 under and postgraduate students in the University of A Coruna (UDC) during October 2013. The experiment will be determinant to assess the goodness of our tests by comparing the results obtained with the regular results in the literature.

Three main contributions of this paper are in order. First, we design two short tests 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 compare 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 insights 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 state of the art, we describe how our tests were designed, firstly on overconfidence and then on prospect theory. Section 3 discusses the results and the reliability of our tests according to the experimental evidence. Section 4 tests some hypotheses about the relationship between demographic priors and behavioral variables. Finally, Section 5 concludes.

2. OVERCONFIDENCE AND PROSPECT THEORY: THEORY AND EXPERIMENT DESIGN

2.1. Overconfidence

The prevalence of overconfidence 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 follow Moore and Healy (2008) for several reasons. First, the three measures of overconfidence have been widely accepted since then (e.g., Glaser et al., 2013). Second, they were able to make a synthesis of the previous debate between the cognitive bias interpretation and the ecological and error models. Third, their model predicts both over and underconfidence in two manifestations (estimation and placement). Fourth, they

ask for frequency judgments across several sets of items of diverse difficulty to account for the evidence that frequency judgments are less prone to display overconfidence, and for the hard-easy effect –a tendency to overconfidence in difficult tasks and underconfidence in easy ones. Finally, their tests are really simple, allowing us to implement an efficient test for overestimation and overplacement that requires only a few minutes to perform it.

Overprecision 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), so in order to avoid confusing them we study overestimation by measuring perceptions across a set of items, while overprecision is analyzed through a series of questions on interval estimates.

Test design

The tests 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** of each respondent.

Our test for **E** and **P** is a simple version of Moore and Healy (2008)'s trivia tests –indeed, several questions were taken from their tests.[1] Participants are required to complete a set of 4 trivia with 10 items each one. To account for the hard-easy effect, two quizzes were easy and two of hard difficulty. 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 with a time limit (150 seconds per trivia) to have a somehow mixed scenario. Prior to solving the trivia, participants were instructed and solved a practice question to familiarize with the experimental setting. Then they took the 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 process for 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

[1] We thank the authors for providing their tests online, they were really helpful to us. We would like to be helpful to other researchers as well: the questions in our tests are available at www.dpeon.com/documentos

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

$$\mathbf{E} = E[X_i] - x_i \quad (1)$$

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

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

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.

Overprecision is analyzed through a separate set of six questions. These tests usually require confidence intervals estimations, but overconfidence in interval estimates may result from variability in setting interval widths (Soll and Klayman, 2004). Hence, in order to disentangle variability and true overprecision, they define the ratio

$$\mathbf{M} = \text{MEAD}/\text{MAD} \quad (3)$$

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. Thus, **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.[3]

Soll and Klayman show that different domains are systematically associated with different degrees of overconfidence and that asking for three fractile estimates rather than two reduces overconfidence. With these results in mind, we devised our test as follows. First, we

[3] Soll and Klayman's methodology also has its flaws: Glaser et al. (2013) discuss the difficulty to compare the width of intervals and different scales and for varying knowledge levels.

ask participants to specify a three-point estimate (median, 10% and 90% fractiles). Second, since we can only ask a few questions and the risks of relying on a single domain were emphasized, we choose to make a pair of questions on three different domains. Thus, questions 1 to 4 are traditional almanac questions on two different domains –the year a device was invented and mortality rates (Shefrin, 2008). Most studies ask judges to draw information only from their knowledge and memory. Soll and Klayman introduce a variation: domains for which participants could draw on direct, personal experience. We do the same in questions 5 and 6 to ask, inspired by Soll and Klayman, about 'time required to walk from one place to another in the city at a moderate (5 km/h) rate'.

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 a ratio *M* for each domain. Consequently, we have 3 different estimations of the ratio. ***M*** could then simply be calculated as either the average ($\mathbf{M}_{avg}$) or the median ($\mathbf{M}_{med}$) of the 3 different estimations.

2.2. Prospect theory

Prospect theory, PT, 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. Thus, assume two mutually 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 $v(\cdot)$ 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:

$$w(p) \cdot v(c_1) + w(1-p) \cdot v(c_2) \ . \qquad (4)$$

The value function has three essential characteristics: reference dependence, diminishing sensitivity and loss aversion. 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 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.

The value and weighting functions suggested by Kahneman and Tversky (1979) are able to explain that fourfold pattern. However, prospect theory as initially defined may lead to a violation of in-betweenness. To avoid this, Tversky and Kahneman (1992) introduced cumulative prospect theory, CPT, which applies the probability weighting to the cumulative distribution function, in a way Eq. (4) becomes

$$w(p) \cdot v(c_1) + 1 - w(p) \cdot v(c_2) \tag{5}$$

for binary lotteries. Yet, Rieger and Wang (2008) observe that not all properties of CPT correspond well with experimental data and that there are some descriptive reasons favoring the original formulation of PT (Hens and Rieger, 2010). The solution they offer allows to generalize prospect theory to non-discrete outcomes and to make it continuous. Their approach is computationally easier than CPT: it simply starts with the original formulation of prospect theory in (4), and fixes the violation of in-betweeness by simply normalizing the decision weights w(p) so that they add up to 1 and can be interpreted again as a probability distribution (Hens and Bachmann, 2008). The approach goes back to Karmakar (1978) where, for two-outcome prospects, the PT-values are normalized by the sum of the weighted probabilities. Thus, the normalized weights $w^*(p)$ are calculated as

$$w^*(p) = \frac{w(p)}{w(p) + w(1-p)} \tag{6}$$

where $w^*(p)$ means normalized weights according to this so-called normalized prospect theory (NPT). NPT has some advantages. Firstly, it cures the violations of state-dominance in lotteries with two outcomes and avoids violations of in-betweenness completely (Hens and Bachmann, 2008). In addition, it is shown that the normalized PT utility converges to a continuous distribution –Rieger and Wang (2008) call the resulting model smooth prospect theory (SPT). Finally, it is an easier approach to compute that, in particular, simplifies the computation of the loss aversion parameter in our questionnaires. Consequently, rather than the cumulative prospect theory –more frequently used in the literature- NPT is the approach 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. We choose two classic specifications. First, the (piecewise) power function by Tversky and Kahneman (1992),

$$v(x) = \begin{cases} x^{\alpha+} & for\, x \geq 0 \\ -\beta(-x)^{\alpha-} & for\, x < 0 \end{cases} \quad (7)$$

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 for 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

$$w(p) = \exp(-(-\log(p))^{\gamma}) \quad (8)$$

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 ($\alpha^+, \gamma^+, \alpha^-, \gamma^-$ and β) to estimate. However, we must deal with the problem with loss aversion: 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 $\alpha^+ = \alpha^-$ loss aversion can be a dimensionless quantity. Alternative interpretations of loss aversion were provided (see Booij et al., 2010, for a discussion), but none of them are a straight index of loss aversion –instead, they formulate it as a property of the utility function over a whole range.

A second dispute is how to measure loss aversion, requiring to determine simulatenously the utility for gains and losses. Some authors provide alternative solutions (see for instance Abdellaoui et al. 2008, Booij et al. 2010), but the debate is still open. We opt for a solution inspired by Booij et al. (2010) –by picking up *"all the questions around the zero outcome"* (p.130)- and by the empirical finding that utility is close to linear for moderate amounts of money (Rabin, 2000). Thus, we ask for a few prospects with small amounts of money and assume $\alpha^+ = \alpha^- = 1$ to estimate β (as either a mean or median across prospects).[4]

[4] 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 of the relevant values of x is not an arbitrary choice (Booij et al., 2010). In addition, Por and Budescu (2013) discuss some violations of the gain-loss separability which may limit the generalization of results from studies of single-domain prospects to mixed prospects.

Test design

For parameter estimation, various elicitation methods have been proposed in the literature. Our method merges some characteristics of Tversky and Kahneman's (1992) approach to elicit certainty equivalents of prospects with just two outcomes and Abdellaoui et al.'s (2008) proposal to make an efficient test with a minimum number of questions. Thus, 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 jointly calibrate α^+ and γ^+ and six questions for negative prospects to calibrate α^- and γ^-, using a nonlinear regression procedure separately for each subject. Finally, three questions regarding the acceptability of mixed prospects, in order to estimate β.

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 $\alpha^+, \gamma^+, \alpha^-$ and γ^- used significant, albeit hypothetical, amounts of money of 500, 1,000 and 2,000 euros –in multiples of 500 euros to facilitate the task (Abdellaoui et al., 2008). Second, only the three questions devised to estimate β used small amounts of money for reasons already described. Since larger amounts might affect the perception of the smaller ones in the β elicitation, these three questions were asked in first order. Finally, prior to solving any trial, respondents answered a practice question. Instructions emphasized there were no right or wrong answers (Booij et al., 2010), but that completing the questionnaire with diligence 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, 10 and 100 euros in three consecutive iterations. 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 makes the utility of a gain (V^+) equivalent to the disutility of a loss (V^-). Hence, for the power value function we have

$$\beta = \frac{G^{\alpha+}}{(-L)^{\alpha-}} \tag{9}$$

where *G* means gains, *L* losses, and loss aversion equals the ratio $G/|L|$ when $\alpha^+ = \alpha^-$ –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 proposed, again in sequential order. Figure 1 shows one of the iterations participants had to answer. Respondents 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 in numerical and graphical form. Every time a subject answered whether she preferred the prospect or the sure gain, a new outcome was provided. The process was repeated until the computer informed the question was completed and she could continue with another prospect. The probabilities of success in all six prospects were different (having two questions with probability of success 50% and one with 99%, 95%, 5% and 1%, respectively), which was emphasized to avoid wrong answers.[5] 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. Then, the certainty equivalent of a prospect was 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. 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. Certainty equivalents were estimated similarly (for values in absolute terms).

3. EXPERIMENTAL RESULTS. GOODNESS OF TEST RESULTS

We organized a series of five experimental sessions during October 2013 in the Faculty of Business and Economics (*University of A Coruna*, UDC). A sample of students of different levels and degrees was selected. To make the call, which was open to the target groups, we

[5] 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.

got in direct contact with students to explain what the experiment would consist of, 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 participant per session would win a prize of 60 euros.[6] In total 126 volunteers, all of them under and postgraduate 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 tests, subjects signed a consent form and completed a questionnaire on demographic information about their (a) gender, (b) age, academic background –about (c) level and (d) degree- and (e) professional experience. Then they completed the tests. In what estimations for the behavioral variables is concerned, Table 1 summarizes the basic univariate statistics.

[Insert Table 1 here]

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

3.1 Reliability of tests on Overconfidence

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

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

First, subjects on average exhibited overestimation (clearly) and underplacement. Thus, the average respondent overestimated her performance by 2.9 right answers (in 40 questions), a bias persistent in both easy and hard tests. Besides, 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 the literature supporting a general bias towards overestimation of our abilities (Lichtenstein et al., 1982; De Bondt

[6] A classic problem of field experiments is in regards of their external validity. Incentives often improve external validity (see Peón et al., 2014). Thus, we incorporated the incentive of a 60 euro prize in the strategy game, while participants were informed that their right to claim the prize was conditioned to their diligent behavior in the behavioral tests. They were also informed 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.

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 our performance relative to others on difficult tasks (Moore and Small, 2007). Table 2 summarizes average responses (out of 10 questions per trivia).

[Insert Table 2 here]

Second, there is a strong correlation between **E** and **P**. That is, though the biases along the sample are towards overestimation and underplacement, participants with the highest overestimation tend to consider themselves above average (or, at least, feature a lower underplacement) and vice versa. This supports the interpretation of overestimation and overplacement as interchangeable manifestations of self-enhancement (Kwan et al., 2004).

Finally, the trivia tests were devised to control for the hard – easy effect, but results suggest we failed to propose proper easy tests. As we may see in Table 2 above, hard trivia tests T2 and T3 had average (median) correct answers of 2.29 (2.0) and 2.75 (3.0) –where 2.0 correct anwers may be attributed, on average, only to good luck.[7] However, easier tests T1 and T4 were expected to yield correct answers of 7.0 to 8.0 on average,[8] but respondents on average (median) only hit the right answer 5.4 (5.0) and 5.58 (6.0) out of 10 questions. This would represent a couple of tests of a medium –rather than 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 (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 (graph on the RHS in Figure 2) meet the mentioned tendency towards overestimation and underplacement. For tests with a medium difficulty (graph on the LHS) the general drift upwards is noticeable (what implies that 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. Moreover, the above-mentioned correlation between overestimation and overprecision exists in both instances.

---

[7] 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, 2.0 right answers out of 10.

[8] 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. Otherwise, readers may also attribute it to researchers' overconfidence.

*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.

These are 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, higher than the 14% overconfidence observed by Soll and Klayman (2004) for three-point estimates but about the same level than for a range estimate. Overconfidence varied across domains as expected: the lowest degree of overprecision corresponds to the domain where participants could draw on personal experience ('time to walk'). 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, 97.6% in the highest, and 97.6% (median) if a single ratio per judge is obtained. Finally, we use Soll and Klayman's alternative refinement to estimate *M* to see[9] overprecision is mainly attributable to narrow size intervals. Table 3 summarizes all these results.

[Insert Table 3 here]

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

Though results on aggregate are consistent with empirical literature, we are concerned about the reliability of data at the individual level for several reasons. First, there is evidence that some participants did not understand the instructions. Incidents include a respondent

[9] The original refinement takes the estimates of MEAD and MAD based on the beta function that better fits the three point estimations by the respondent. Alternatively, Soll and Klayman (2004) suggest to measure MAD assuming the median is in the middle of the distribution. They denote M_3 the first ratio and M_2 the second one.

with missing responses, minimum and maximum boundaries swapped, answers provided in a different order than required, and median estimations identical to a boundary.[10] In future research, we suggest to ask participants to fill some boxes in the order lower bound – median – upper bound. According to Soll and Klayman (2004), "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.

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 4.

[Insert Table 4 here]

First, the last refinement method that assumes normality yields the most extreme results. We will see this effect is not a problem of this method but an evidence of a weakness of the test. Second, indicators computed as average ratios are higher. Third, if we compare how many individuals have an estimator that varies substantially[11] whether we use medians or averages, about half of the individuals have a sensible indicator. This effect is particularly pernicious when the *M* ratios yield conflicting qualitative results: that is, a same individual with overprecision (M<1) or underprecision (M>1) depending on the method used. 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[12] (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: six questions revealed 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 $M \to \infty$, which

[10] Fortunately, we could contact participants after the experiment to confirm their answers. Thus, errors like swapped boundaries or responses in a particular order could be amended. Others, like missing values or median estimations equal to a boundary, were not modified as it would represent an alteration of the experiment results.
[11] We consider a 'substantial variation' of 25% between median and average estimations: for median estimations of *M* about 0.40, this makes 0.10 in absolute terms. This variation is equivalent to the median variations observed along the sample for the three refinement methods (0.09 – 0.10 according to Table 4).
[12] This comparison across methods analyzes the minimum and maximum estimations for each individual using any of the three methods.

distorts 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 medians. Though this effect is more palpable for the refinement method that assumes normality, this only happened by chance. In particular, there were a few respondents 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 indicator we are using as option-by-default.

3.2 Reliability of tests on Prospect Theory

Respondents completed the fifteen questions in about 20 minutes, instructions included, and there were no relevant incidents in any of the five sessions. 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

Table 5 provides the results at the aggregate level, which 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 respondent. We also compare our results against some classic results in the literature (where the power and Prelec-I functions were used).[13]

[Insert Table 5 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., 2010). 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^-).

[13] 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. (2010). 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. (2010).

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 $\gamma^- = 0.5$) in consonance with literature. Using Prelec-I function we are imposing the classic inverse S-shaped weighting function (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.[14]

Parameters α^- and γ^- suggest a strong risk seeking behavior in the negative domain. There may be two interpretations not mutually exclusive. First, most participants were unable to fully interpret hypothetical losses as real: several participants were strongly biased in terms of probability weighting (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 profiles might be better described with a weighting function that accounts for elevation as well as curvature, like Prelec-II. Besides, a contamination effect might also affect α^- estimations.

Finally, our beta estimations are in consonance with classic results (a loss aversion higher than 2) compared to the more moderate estimations reported by Booij et al. (2010). 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 given 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

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

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 6.[15] 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 6 here]

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 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 repeat two types of iterations[16] to obtain 96% reliability for the first replication and 66% for the second one, and claim them to be satisfactory. Following this, we repeat 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: only 5.6% of responses were contradictory (94.4% reliability). Furthermore, 65.4% of participants made not any response error, 81.7% made one at most, and only 2 out of 126 participants made more than three. This confirms the experiment design was helpful for participants to correctly understand the task. Whether some risk profiles are not common may hence be attributed to the difficulties for some participants to imagine hypothetical losses as real, but not to a misinterpretation of data.

Regarding fitting errors, the high quality of the R^2 coefficients obtained to estimate the parameters for most individuals are both a confirmation that participants understood the

[15] For risk-neutrality in Table 6 we report the percentage of elicitations that revealed a certainty equivalent that was the closest possible to the expected value of the game.
[16] 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.

task, and 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 the section 'anomalies at the individual level' below we analyze some results that might reveal some mistakes or confusion by the respondent. Table 7 summarizes the R^2 obtained.

[Insert Table 7 here]

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[17] whose risk attitudes 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.

[17] These subjects show the lowest fitting accuracy in any of the two domains or both. In addition, it seems these judges had more problems to understand the task: all but one made at least one iteration error, an average of 1.75 errors per respondent statistically higher ($p < 0.01$) than the 0.61 mean error of all the other participants.

4. HYPOTHESIS TESTING

Once the validity of the questionnaires we devised has been confirmed, we now use the estimated measures to test two types of hypotheses: the effect of some priors over the behavioral variables, and the relationship among variables. This way, we aim to contribute to the behavioral literature in two instances: providing additional data on the relationship between behavioral biases and priors such as gender, age or experience on one hand, and the relationship between all measures of overconfidence and risk profile according to prospect theory on the other. In particular, the experimental analysis of prospect theory and overconfidence for a same sample of participants is something that, to the best of our knowledge, has not been done before.

Table 8 summarizes the priors on demographic information subjects were required to provide about gender, age, academic year (level), degree and professional experience, as well as the values they may take. The descriptive statistics of age and level, as well as of the behavioral biases, were provided in Table 1.

[Insert Table 8 here]

Univariate analysis highlights some pros and cons of our sample. On the negative side, all participants are college students. Consequently, the sample is limited in terms of age (98.4% of respondents were between 17 and 28 years old) and level happens to be not a good proxy for education: for hypothesis testing in the literature, education is intended to 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 (as well as with working experience). This will represent a drawback for the hypothesis testing below. 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 included a subsample of 21 students that have no degree in economic or financial studies to serve as contrast.

We pose several hypothesis, based on extensive literature review. On one hand, in regards to the effect of priors over the prospect theory and overconfidence measures, we test the following set. 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 overconfidence (Sandroni and Squintani, 2009; Zell and Alicke, 2011; while Hanson et al., 2008 suggest it increases). 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.

On the other hand, we trace relationships among variables of three kinds: among different overconfidence measures, among prospect theory parameters, and between prospect theory and overconfidence. 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}) –denoted age (r) and βavg (r) in Table 9.

[Insert Table 9 here]

We test the hypotheses with a correlation and a regression analysis. The most relevant results follow in order. Regarding priors and variables, we find a significant correlation 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. We also find experience reduces objectivity in terms of estimation of self-performance ($p < 0.05$) –contrary to hypothesis (vii). The hypotheses on gender and skills were tested with an ANOVA test. Regarding gender, women appear to be more overconfident 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. The first result supports hypothesis (xi) while the second one goes against (x).[19]

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's (2008) assertion that overprecision reduces them both.

[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 correlations between priors and variables go in the same direction as the null hypotheses tested, but with no statistical significance. First, age reduces overconfidence: older students exhibit lower levels of overestimation and overplacement (with no significance) as well as 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. Fourth, 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. Fifth, regarding skills in finance, it reduces overestimation and increases loss aversion (β_{med}).

Correlation among PT measures 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. Finally, in regards to the relationship between overconfidence and PT parameters, we only find positive correlations ($p < 10\%$) between α^- and **E,** and between γ^- and **M**. These 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). However, this result might also be consistent with Kahneman and Lovallo's (1993) suggestion that biases can cancel out.

Finally, the regression analysis yields results that are coherent with correlations. Thus, we regress 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 to validate our tests confirm they 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 unable 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 paper contributes with a set of short tests that are able to obtain efficient results overall. Besides, it provides additional evidence about how gender, education and skills in finance affect overconfidence and risk aversion. In particular, our analysis enhances the scope for empirical application of prospect theory and overconfidence by using the same sample of respondents in the experimental analysis –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. In addition, it would be interesting to extend the way we tested the behavioral biases of the participants to other tests and elicitation methods available, such as cumulative prospect theory, non-parametric methods, and others. Two open questions in the test for prospect theory 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., 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.

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.

Glaser, M., T. Langer and M. Weber (2013), True overconfidence in interval estimates: Evidence based on a new measure of miscalibration, *Journal of Behavioral Decision Making* 26, 405-417.

Hanson, P., M. Ronnlund, P. Juslin and L.G. Nilsson (2008), Adult age differences in the realism of confidence judgments: Overconfidence, format dependence, and cognitive predictors, *Psychology and Aging* 23(3), 531-544.

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

Hens, T. and M.O. Rieger (2010), *Financial economics: A concise introduction to classical and behavioral finance*, Springer.

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.

Kahneman, D. and D. Lovallo (1993), Timid choices and bold forecasts: A cognitive perspective on risk taking, *Management Science* 39(1), 17-31.

Karmakar, U.S. (1978), Subjectively weighted utility: A descriptive extension of the expected utility model, *Organizational Behavior and Human Performance* 24, 67-72.

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.

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.

Por, H.-H. and D.V. Budescu (2013), Revisiting the gain–loss separability assumption in prospect theory, *Journal of Behavioral Decision Making* 26, 385–396.

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.

Rieger, M.O. and M. Wang (2008), Prospect theory for continuous distributions, *Journal of Risk and Uncertainty* 36, 83–102.

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.

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

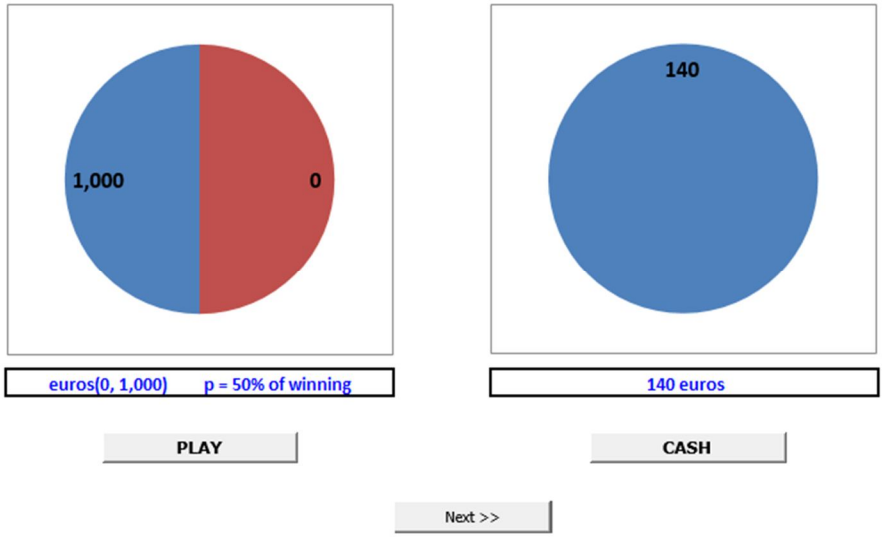


FIGURE 2 – The hard – easy effect

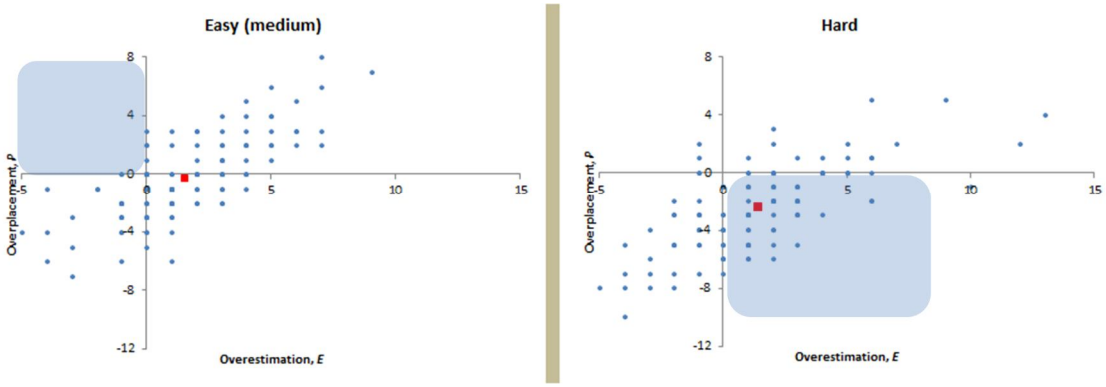
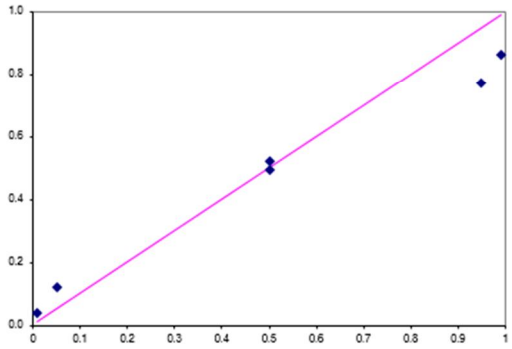


FIGURE 3 – Risk attitudes of the average participant

	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
0.01	0.05	0.5	0.5	0.95	0.99
0.018	0.063	0.267	0.220	0.620	0.691
risk averse	risk averse	risk seeking	risk seeking	risk seeking	risk seeking

Positive prospects



Negative prospects

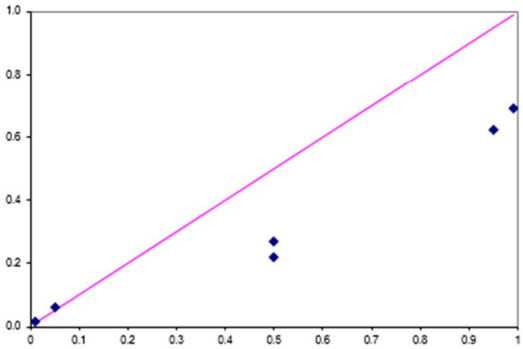


FIGURE 4 – Risk attitudes of eight individual anomalies

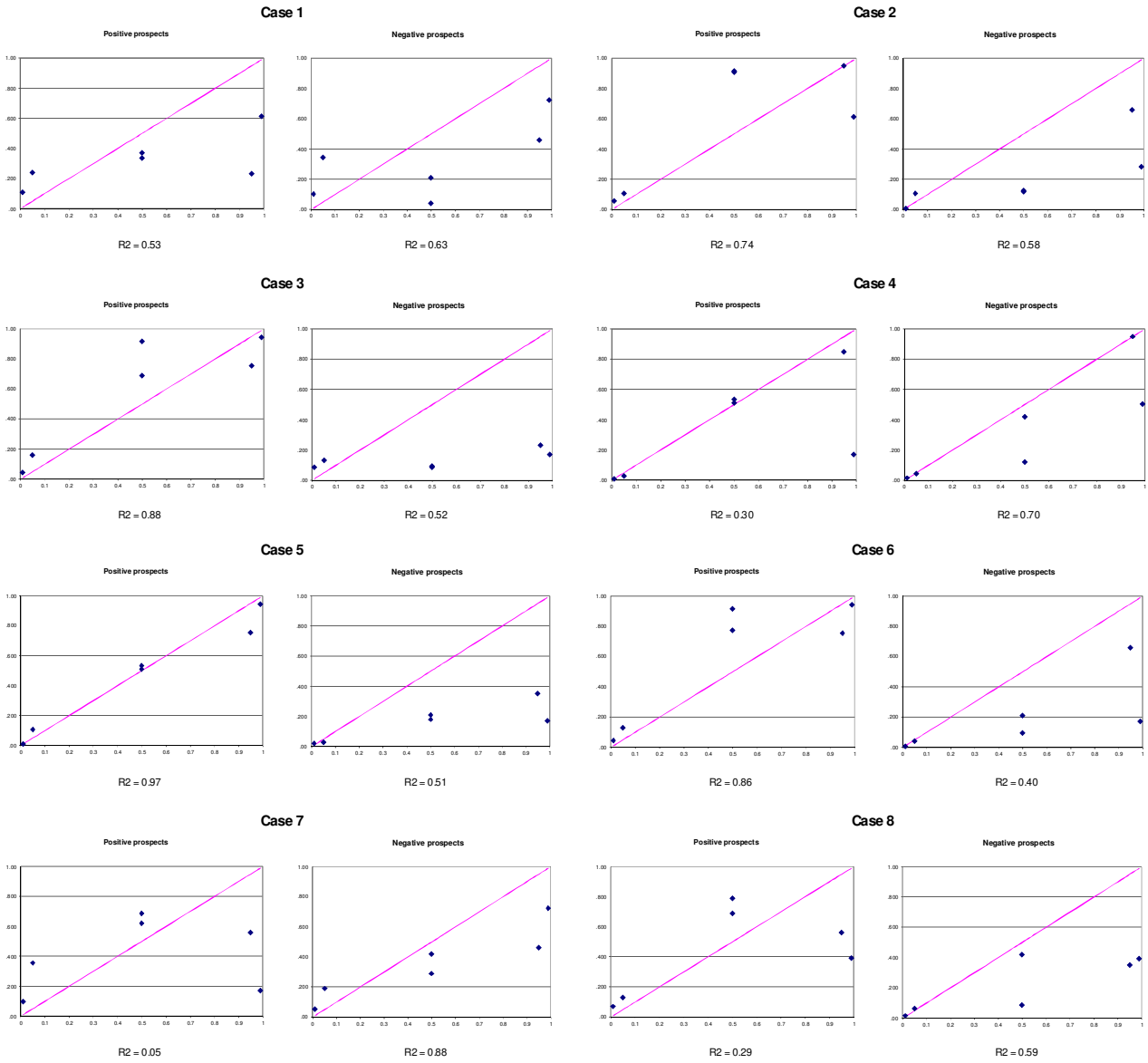


TABLE 1 – Descriptive statistics of behavioral variables

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

Measures M_{med} and M_{avg} have 125 observations due to missing responses by one participant.

TABLE 2 – 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 3 – 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 4 – 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

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

across meth.	median	average
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 5 – 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'10: $\alpha^+ = 0.86$
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'10: $\alpha^- = 0.83$
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'10 review: 1.38 to 1.63 - Booij'10 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'10 (Booij et al., 2010)

** These research articles imposed a different parametric specifications other than Prelec-I for the weighting function

TABLE 6 – 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%

	GAINS		LOSSES	
	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 7 – 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%

TABLE 8 – 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 10 – Summary of priors

Correlations

		Age (r)	Level	Exper.	E	P	Mmed	Mavg	alpha +	alpha -	gamma +	gamma -	βmed	βavg (r)
Age (r)	Pearson Correlation	1	,616**	,403**	-.030	-.065	.111	.114	.070	.033	.056	-.035	.042	.090
	Sig. (2-tailed)		.000	.000	.738	.474	.221	.208	.439	.714	.532	.699	.643	.324
	N	125	125	125	125	125	124	124	125	125	125	125	125	122
Level	Pearson Correlation	,616**	1	,210*	-.024	.047	-.027	.041	-.001	-.029	.018	-.054	.174	,209*
	Sig. (2-tailed)	.000		.018	.790	.598	.764	.647	.987	.746	.842	.551	.051	.020
	N	125	126	126	126	126	125	125	126	126	126	126	126	123
Experience	Pearson Correlation	,403**	,210*	1	.151	.046	.035	-.060	.077	-.094	.054	-.105	-.036	-.002
	Sig. (2-tailed)	.000	.018		.091	.611	.700	.508	.394	.294	.548	.244	.690	.980
	N	125	126	126	126	126	125	125	126	126	126	126	126	123
E	Pearson Correlation	-.030	-.024	.151	1	,690**	-.123	-.166	-.055	.165	.055	-.025	-.065	-,199*
	Sig. (2-tailed)	.738	.790	.091		.000	.173	.064	.544	.065	.537	.782	.468	.027
	N	125	126	126	126	126	125	125	126	126	126	126	126	123
P	Pearson Correlation	-.065	.047	.046	,690**	1	-.039	-.144	-.122	.096	.104	.069	-.010	-.089
	Sig. (2-tailed)	.474	.598	.611	.000		.664	.109	.174	.285	.246	.441	.913	.325
	N	125	126	126	126	126	125	125	126	126	126	126	126	123
Mmed	Pearson Correlation	.111	-.027	.035	-.123	-.039	1	,672**	-.121	.008	.032	.165	.052	.087
	Sig. (2-tailed)	.221	.764	.700	.173	.664		.000	.180	.932	.723	.066	.564	.339
	N	124	125	125	125	125	125	125	125	125	125	125	125	122
Mavg	Pearson Correlation	.114	.041	-.060	-.166	-.144	,672**	1	-.095	.041	-.021	.144	.122	,193*
	Sig. (2-tailed)	.208	.647	.508	.064	.109	.000		.293	.653	.812	.110	.175	.033
	N	124	125	125	125	125	125	125	125	125	125	125	125	122
alpha +	Pearson Correlation	.070	-.001	.077	-.055	-.122	-.121	-.095	1	-,211*	,597**	-.133	-.036	-.040
	Sig. (2-tailed)	.439	.987	.394	.544	.174	.180	.293		.018	.000	.139	.687	.658
	N	125	126	126	126	126	125	125	126	126	126	126	126	123
alpha -	Pearson Correlation	.033	-.029	-.094	.165	.096	.008	.041	-,211*	1	-.093	,326**	,294**	,308**
	Sig. (2-tailed)	.714	.746	.294	.065	.285	.932	.653	.018		.301	.000	.001	.001
	N	125	126	126	126	126	125	125	126	126	126	126	126	123
gamma +	Pearson Correlation	.056	.018	.054	.055	.104	.032	-.021	,597**	-.093	1	,250**	-.068	-.097
	Sig. (2-tailed)	.532	.842	.548	.537	.246	.723	.812	.000	.301		.005	.450	.287
	N	125	126	126	126	126	125	125	126	126	126	126	126	123
gamma -	Pearson Correlation	-.035	-.054	-.105	-.025	.069	.165	.144	-.133	,326**	,250**	1	-.035	.008
	Sig. (2-tailed)	.699	.551	.244	.782	.441	.066	.110	.139	.000	.005		.698	.926
	N	125	126	126	126	126	125	125	126	126	126	126	126	123
βmed	Pearson Correlation	.042	.174	-.036	-.065	-.010	.052	.122	-.036	,294**	-.068	-.035	1	,918**
	Sig. (2-tailed)	.643	.051	.690	.468	.913	.564	.175	.687	.001	.450	.698		.000
	N	125	126	126	126	126	125	125	126	126	126	126	126	123
βavg (r)	Pearson Correlation	.090	,209*	-.002	-,199*	-.089	.087	,193*	-.040	,308**	-.097	.008	,918**	1
	Sig. (2-tailed)	.324	.020	.980	.027	.325	.339	.033	.658	.001	.287	.926	.000	
	N	122	123	123	123	123	122	122	123	123	123	123	123	123

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).